UNIVERSIDADE FEDERAL FLUMINENSE

RENATHA OLIVA CAPUA

Identificação Estatística de Regiões Codificadoras de Proteínas em Seqüências de DNA

NITERÓI

2005

UNIVERSIDADE FEDERAL FLUMINENSE

RENATHA OLIVA CAPUA

Identificação Estatística de Regiões Codificadoras de Proteínas em Seqüências de DNA

Dissertação de Mestrado submetida ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do título de Mestre em Computação. Área de concentração: Otimização Combinatória e Inteligência Artificial.

Orientadora:
Helena Cristina da Gama Leitão

Co-orientador:
Jorge Stolfi

NITERÓI

2005

Identificação Estatística de Regiões Codificadoras de Proteínas em Seqüências de DNA

Renatha Oliva Capua

Dissertação de Mestrado submetida ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do título de Mestre em Computação.

Aprovada por:

Profa. Helena Cristina da Gama Leitão, D.Sc. / IC-UFF
(Presidente)

Prof. Jorge Stolfi, Ph.D./ IC-UNICAMP

Profa. Celina Miraglia Herrera de Figueiredo, D.Sc./
COPPE-UFRJ

Prof. João Meidanis, Ph.D./ IC-UNICAMP

Prof. Alexandre Plastino de Carvalho, D.Sc. / IC-UFF

Niterói - RJ, 21 de Novembro de 2005.

À mamãe e ao papai.

Agradecimentos

Agradeço em primeiro lugar a Deus, meu guia nesta caminhada e que me ajudou a ter força para enfrentar os desafios.

Obrigada à mamãe e ao papai, pessoas mais importantes na minha vida, por me financiarem, incentivarem, por acreditarem em mim e na minha capacidade e por não medirem esforços para me ver feliz.

Agradeço à Profa. Helena pela orientação recebida, pela amizade, e por todos os lanchinhos. Meu muito obrigada por toda a atenção dispensada a mim e sem a qual não teria realizado este trabalho. Agradeço também ao Prof. Stolfi pela orientação e pelas grandes e valiosas sugestões.

Aos professores Cristina Boeres e Marcone Jamilson Freitas, pela força e incentivos recebidos durante e antes do mestrado; aos professores Satoru e Otton, pela amizade.

Agradeço aos amigos que me acompanharam durante o mestrado. Às grandes Dani, Vivi e Cris que me acolheram no nosso doce lar, anexo 1 do laboratório. Aos numerosos amigos com quem morei, em especial ao Glauco (Grauco), grande confidente, e ao Alex. Aos inesquecíveis amigos de todas as horas, cafezinhos e cantareiras: Ivairton (Vairtu), Stênio (InSAno), Ju, Lu Brugiolo, Dani, Vivi, Jacquinho, Rodrigo, Jonivan, Lu Pessoa e Bruno. À Tânia, ao Antônio, à Sandra e à Heloisa, que me presentearam com a sua amizade no catecumenato. Aos amigos Robertinha, Mê e Mariozinho, que mesmo de longe souberam estar presentes. Ao Ary e Leandro Ciuffo, que me acompanharam nos momentos desesperadores das disciplinas. Obrigado a Luiz Merschmann, Idalmis, Adria, Tiago jovem, Tiago faquinha, Deolinda, Mokado, Cris Maciel, Luciene Motta, Haroldo, Maysa, Dudu Corrêa, Adriana Becharra, Luciano, Kennedy e muitos outros que me ajudaram de alguma forma.

Obrigada ao Marcel, pela companhia nas madrugadas de estudo no laboratório, pela correção do texto e principalmente pela persistência.

Agradeço aos membros da banca, pelas sugestões recebidas. E às secretarias da pós Ângela, Izabela e Maria, pela grande ajuda e amizade.

Resumo

O elevado número de projetos de seqüenciamento de genomas em andamento e a conseqüente geração de grandes quantidades de dados descaracterizados tem motivado a busca por métodos computacionais mais precisos e eficientes para a identificação das estruturas que compõem o DNA dos seres vivos. Em especial, devido a sua grande importância, destaca-se a busca por regiões codificadoras de proteínas, que vem sendo o foco de pesquisas há pelo menos vinte anos. Estas regiões armazenam em seus nucleotídeos a informação necessária às estruturas celulares para a fabricação das proteínas, componente fundamental da maioria dos organismos vivos.

A identificação das regiões codificadoras nas seqüências de DNA ainda é um problema de difícil solução, uma vez que os complexos mecanismos celulares envolvidos no processo de fabricação das proteínas não são completamente conhecidos.

Neste trabalho, desenvolvemos um método estatístico para a identificação das regiões codificadoras de proteínas. O método é baseado no teorema de Bayes aplicado a trechos de k bases consecutivas do DNA, onde k é um parâmetro especificado pelo usuário. Para o cálculo das probabilidades condicionais e *a priori* necessárias para o teorema de Bayes, usamos certas hipóteses sobre independência de bases e códons, e sobre o tamanho mínimo de regiões codificadoras e não-codificadoras, que reduzem o custo computacional e o tamanho das tabelas de probabilidade. Em testes realizados, o método proposto apresentou resultados promissores.

Palavras-chave: Seqüências codificadoras de proteínas, teorema de Bayes, reconhecimento de padrões, bioinformática.

Abstract

The large number of genome sequencing projects in progress and the resulting increase in the volume of uncharacterized data has motivated the search for more precise and efficient computacional methods for identifying the structures that compose the DNA of living beings. In particular, due to its great importance, the search for protein coding regions has been the focus of research for at least twenty years. Coding regions carry in its nucleotides the information necessary to the cellular structures to produce proteins, fundamental component of most living organisms.

The identification of coding regions in DNA sequences is still a difficult problem since the complex cellular mechanisms involved in the process of protein production are not completely known.

In this dissertation, we have developed a statistical method for the identification of protein coding regions. The method is based on Bayes's theorem applied to strings of k consecutive DNA bases, where k is a parameter specified by the user. To compute the conditional and a priori probabilities needed by Bayes's theorem, we use certain hypotheses on the independence of codons and bases, and on the minimum size of coding and non-coding regions, that reduce the computational cost and the size of probability tables. In performed tests the proposed method has presented promising results.

Keywords: Protein coding sequences, Bayes's theorem, pattern recognition, bioinformatics.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Regiões codificadoras e não-codificadoras

A Biologia Computacional é a área da Ciência da Computação que busca resolver problemas envolvendo bioseqüências — DNA, RNA e *proteínas*. Destas, as proteínas são encarregadas de manter e controlar os processos vitais nos seres vivos, ou seja, são responsáveis pela existência da vida no sentido físico [14]. As informações necessárias à produção de proteínas estão armazenadas em trechos das seqüências de *DNA* e *RNA*, chamadas *codificadoras*, que são interrompidas por seqüências de bases que não são usadas na produção de proteínas, chamadas de *não-codificadoras*.

A identificação de regiões codificadoras é um dos problemas fundamentais da biologia e bioinformática. A busca por estas regiões em genomas, especialmente dos *eucariotos*(animais, plantas e fungos), ainda é um desafio para os cientistas, uma vez que os mecanismos celulares envolvidos na produção das proteínas nestes organismos ainda não são completamente conhecidos. A complexidade da célula *eucariótica*, contendo uma grande quantidade de material genético com função ainda desconhecida, dificulta o reconhecimento das regiões codificadoras.

Hoje em dia, determinar a seqüência de bases do DNA de um ser vivo é uma tarefa trabalhosa mas essencialmente mecânica. Nos últimos anos, com a grande quantidade de projetos de seqüenciamento em andamento, o número de organismos seqüenciados cresceu drasticamente. As seqüências armazenadas no Genbank [2] passaram de aproximadamente 215 mil em 1994 para 41 milhões em 2004 [3], gerando uma enorme quantidade de dados brutos, aguardando interpretação. Desta forma, tornou-se necessária a automatização do processo de reconhecimento das regiões codificadoras, e o desenvolvimento de métodos computacionais eficientes e confiáveis. Porém, ainda não existem programas que resolvam este problema de maneira

completamente satisfatória [17].

1.2 Contribuição deste trabalho

Neste trabalho, propomos um método computacional para resolver o problema do reconhecimento de regiões codificadoras em seqüências de DNA, utilizando as características estatísticas de tais regiões. Métodos baseados nestas características mostram-se mais confiáveis do que os que utilizam exclusivamente os sinais de transcrição ou ínicio de translação, usados pelos mecanismos celulares, por dois fatores principais. O primeiro é devido ao fato de que esses sinais ainda são de difícil caracterização. O segundo é que tais sinais podem não estar presentes em uma seqüência de DNA escolhida para ser analisada [7].

Nosso método calcula, dada uma seqüência de DNA, as probabilidades de cada base dessa seqüência pertencer a uma região codificadora. Para isto, usamos o fato de que as regiões codificadoras possuem propriedades estatísticas diferentes das regiões não-codificadoras, em decorrência do fato de que elas especificam a produção de proteínas segundo um código determinado. Graças a essas diferenças, analisando um trecho de k nucleotídeos consecutivos, podemos deduzir (através do teorema de Bayes) a probabilidade em questão.

Como veremos, esta abordagem também fornece informação que pode ser usada para predizer a seqüência dos aminoácidos da proteína codificadora.

Podemos fazer uma analogia entre este problema e o reconhecimento de um idioma: supondo que temos um texto que, sabidamente, consiste de trechos em português misturados com trechos em inglês, devemos identificar quais trechos pertencem a cada língua. Intuitivamente, quanto maior o trecho analisado, mais fácil distinguir a qual idioma ele pertence. A análise de trechos curtos, como por exemplo somente uma letra, não fornece uma classificação muito confiável, mas a confiabilidade é bem maior quando analisamos 10-20 letras consecutivas.

Da mesma forma, quando analisamos trechos maiores de DNA, podemos distinguir mais claramente as regiões codificadoras das não-codificadoras. Porém, o custo computacional dessa análise aumenta significativamente à medida que aumenta o tamanho dessa “janela” analisada. A principal contribuição deste trabalho é um conjunto de técnicas que permitem diminuir bastante esse custo.

1.3 Apresentação dos resultados

Idealmente, uma ferramenta de classificação deve rotular cada base de um trecho do DNA estudado com marcas “C” ou “N”, indicando respectivamente se a base pertence a uma região codificadora ou não-codificadora. No entanto, no caso da classificação por métodos estatísticos, tal resposta ocasiona perda de informação, uma vez que os resultados obtidos sob a forma de probabilidades devem ser convertidos em uma resposta sim ou não (1 ou 0). Afinal, uma classificação categórica só pode ser conseguida por métodos experimentais, que utilizam técnicas biológicas.

1.4 Estrutura da dissertação

O restante desta dissertação está dividida como segue. No capítulo 2, fazemos uma breve descrição de alguns conceitos básicos de biológia molecular e no capítulo 3, falamos de probabilidade e estatística, apresentando o teorema de Bayes. No capítulo 4, definimos mais formalmente o problema abordado e os trabalhos existentes na literatura. O método proposto é apresentado no capítulo 5 e os resultados encontrados são discutidos no capítulo 6. Finalmente, no capítulo 7 apresentamos as conclusões e as propostas de trabalhos futuros.

Capítulo 2

Conceitos básicos de biologia molecular

Neste capítulo apresentamos breves conceitos de Biologia Molecular necessários ao entendimento do problema em questão. Iniciamos na seção 2.1 por uma descrição dos vírus, dos procariotos e dos eucariotos, ressaltando algumas de suas principais características e diferenças celulares. Na seção 2.2, falaremos sobre as proteínas, componente fundamental dos seres vivos, e nas seções 2.3 e 2.4 apresentamos alguns componentes celulares envolvidos na sua produção. Finalmente, na seção 2.5 explicamos como estes componentes realizam a produção protéica. Este capítulo foi escrito tomando-se como principais referências os textos de Junqueira e Carneiro [11] e Alberts *et al.* [4].

2.1 Vírus, Eucariotos e Procariotos

Os seres vivos são estruturas formadas por várias substâncias orgânicas, principalmente proteínas, *carboidratos*, *lipídios* e *ácidos nucléicos*. Como veremos mais adiante, as proteínas são responsáveis pela maioria das reações químicas que sustentam a vida, e os ácidos nucléicos carregam a informação que determina a estrutura dos seres vivos e controlam o desenvolvimento, funcionamento e reprodução dos organismos.

Os seres vivos podem ser divididos em três grandes grupos: os *vírus*, os *procariotos* e os *eucariotos*. Os vírus são basicamente formados de ácidos nucléicos com um envelope simples de proteínas. Eles não tem vida independente e precisam parasitar outros organismos para se reproduzir. Os procariotos são todas as bactérias, incluindo as algas azuis (cianobactérias). Os eucariotos compreendem todos os demais seres vivos — animais, fungos, protozoários e plantas.

Os procariotos e eucariotos são formados de unidades relativamente auto-contidas chamadas

células, capazes de crescer e se multiplicar por divisão. Toda célula possui uma *membrana* de substâncias gordurosas (fosfolipídios) envolvendo um *citoplasma* composto principalmente de água e proteínas.

Um procarioto é formado por uma única célula (*procarionte*) geralmente pequena e com estrutura interna simples, cujo material genético (*ácidos nucléicos*) está espalhado por todo o citoplasma.

As células de um eucarioto (*eucariontes*) tem estrutura interna mais complexa. Seu citoplasma inclui uma rede relativamente rígida, o *citoesqueleto*, responsável pelos movimentos e pela forma das células. Imersas no citoplasma, há também várias estruturas internas distintas, as *organelas*: o *retículo endoplasmático* (*liso e rugoso*), os *lisossomas*, as *mitocôndrias* e, nas células vegetais, os *cloroplastos*. Dentre as organelas destaca-se o *núcleo*, delimitado pela *membrana nuclear*.

2.2 Proteínas

As proteínas são macromoléculas que compõem a maior parte do nosso corpo e que desempenham um grande conjunto de funções vitais. Incluem por exemplo, as *enzimas*, catalisadores que aceleram as reações químicas; componentes estruturais, como o *colágeno*; componentes regulatórios, como certos *hormônios*; e componentes protetores, como *fatores de coagulação do sangue*.

Estruturalmente, as proteínas são longas cadeias de aminoácidos unidos uns aos outros. Cada aminoácido por sua vez é formado por um átomo de carbono central (chamado de *carbono-*α) ligado a um grupo carboxílico ($\mathrm{COO^-}$), um grupo amino ($\mathrm{NH_3^+}$), e uma cadeia lateral (que determina sua função e diferencia um aminoácido de outro). Esta estrutura é mostrada na figura 2.1. Embora existam muitos aminoácidos possíveis, somente 20 deles (enumerados na tabela 2.1) são normalmente encontrados nos seres vivos.

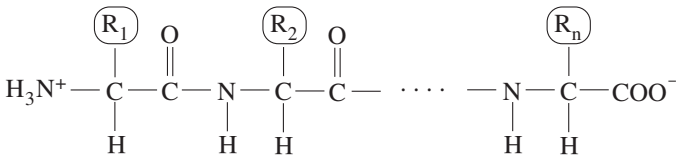

Figura 2.1: Estrutura química de uma proteína.

Dada a grande importância das proteínas, qualquer alteração em sua estrutura pode comprometer de algum modo o funcionamento do organismo. Por exemplo, a simples troca de

Abreviação		Aminoácido
A	Ala	Alanina
C	Cis	Cisteína
D	Asp	Ácido Aspártico
E	Glu	Ácido Glutâmico
F	Fen	Fenilalanina
G	Gli	Glicina
H	His	Histidina
I	Ile	Isoleucina
K	Lis	Lisina
L	Leu	Leucina
M	Met	Metionina
N	Asn	Asparagina
P	Pro	Prolina
Q	Gln	Glutamina
R	Arg	Arginina
S	Ser	Serina
T	Tre	Treonina
V	Val	Valina
W	Trp	Triptofano
Y	Tir	Tirosina

Tabela 2.1: Os 20 aminoácidos mais comumente encontrados nas proteínas.

um determinado aminoácido na hemoglobina humana altera a forma das hemácias (células sangüíneas), resultando em uma doença chamada *anemia falciforme*.

2.3 Ácidos nucléicos

Todos os processos bioquímicos são controlados pelos chamados ácidos nucléicos [4]. Os ácidos nucléicos são longas cadeias formadas pela união de pequenas moléculas chamadas de *nucleotídeos*. Cada nucleotídeo, por sua vez, é constituído por uma base que contém nitrogênio, um açúcar de 5 carbonos (*pentose*) e um grupo fosfato. A estrutura esquemática de uma cadeia de ácido nucléico é exibida na figura 2.2.

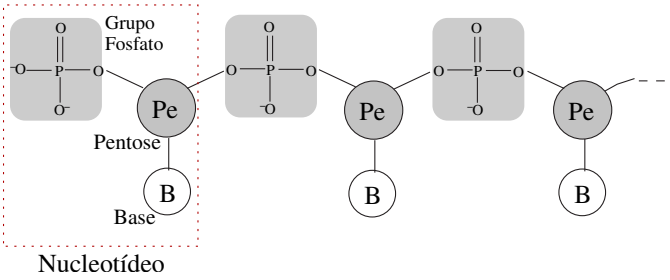

Figura 2.2: Estrutura de uma molécula de um ácido nucléico.

Existem dois tipos de ácidos nucléicos classificados de acordo com a pentose que constitui

seus nucleotídeos: o *ácido ribonucléico*, ou RNA, possui como pentose a *ribose* e o *ácido desoxirribonucléico*, ou DNA, possui a *desoxirribose* (figura 2.3).

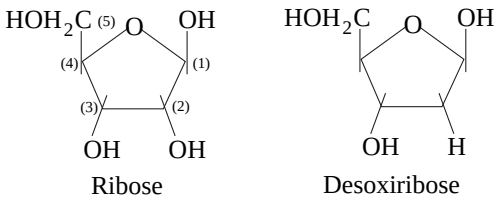

Figura 2.3: Estrutura química de uma ribose e de uma desoxirribose.

Como visto na figura 2.3, os carbonos das pentoses são numerados convencionalmente de 1 a 5. Numa cadeia de DNA ou RNA, o carbono número 5 no açúcar de um nucleotídeo está ligado através do grupo fosfato ao carbono número 3 no açúcar do nucleotídeo seguinte. Desta forma, as cadeias de RNA e DNA possuem uma orientação definida: há uma extremidade com o carbono número 5 livre, chamada extremidade $5'$, e outra, com o carbono número 3 livre, chamada extremidade $3'$.

Cada nucleotídeo é identificado pelo nome de sua base nitrogenada. Com raríssimas exceções, as bases encontradas nos seres vivos são as purinas *adenina* (A) e *guanina* (G), e as pirimidinas *timina* (T), *citosina* (C) e *uracila* (U), vistas na figura 2.4. O DNA é formado pelas bases A, G, C e T, enquanto que o RNA usa A, G, C e U. A seqüência de nucleotídeos ao longo da cadeia armazena a informação necessária para a fabricação das proteínas, como veremos na seção 2.5.

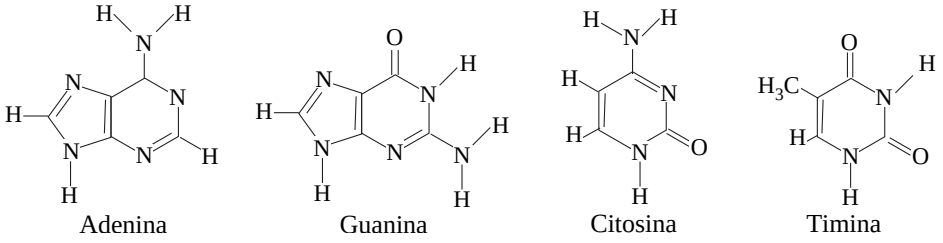

Figura 2.4: Estrutura química das bases adenina, guanina, citosina e timina.

2.3.1 RNA

Em alguns vírus (retrovírus) o RNA é o repositório principal da informação genética. Nos eucariotos e procariotos, entretanto, a principal função do RNA é transportar a informação sobre a estrutura das proteínas do DNA para os mecanismos da célula que realizam a síntese das mesmas. O RNA que desempenha esta tarefa é chamado de *RNA mensageiro*, ou mRNA. Além deste, existe o *RNA ribossômico*, ou rRNA, que junto com outras proteínas, forma o

ribossomo, estrutura que participa da montagem da cadeia de aminoácidos a partir do mRNA. Existe também o *RNA transportador*, ou tRNA, que tem a função de trazer ao ribossomo os aminoácidos por ele requisitados. Tais funcionalidades serão melhor entendidas nas próximas seções.

2.3.2 DNA

Na maioria dos seres vivos, o DNA é responsável pelo armazenamento e transmissão da informação genética através de sua seqüência de nucleotídeos, sendo a base da hereditariedade. É encontrado em certos vírus, e em várias organelas de eucariotos e procariotos, como as mitocôndrias, e os cloroplastos; e, em maior quantidade, dentro do núcleo nos eucariotos, e disperso no citoplasma nos procariotos.

O DNA não é traduzido diretamente para proteínas. Quando o organismo necessita de uma determinada proteína, o trecho de DNA correspondente é transcrito para RNA, que é então traduzido em uma proteína.

Ao contrário do RNA, as cadeias de DNA quase sempre se encontram aos pares. O modelo da molécula de DNA, proposto por Watson e Crick em 1953, é uma dupla hélice composta por duas cadeias de seqüências de nucleotídeos, denominadas *fitas*. As duas fitas são antiparalelas, uma apresenta orientação $5'$ para $3'$ e a outra $3'$ para $5'$. Os nucleotídeos situam-se na parte interna da dupla-hélice; as fitas são ligadas entre si por ligações fracas entre os respectivos nucleotídeos, conhecidas como *pontes de hidrogênio*, formando pares. Cada base timina pareia-se somente com uma base adenina, e cada base citosina pareia-se somente com uma base guanina. Um exemplo pode ser visto na figura 2.5.

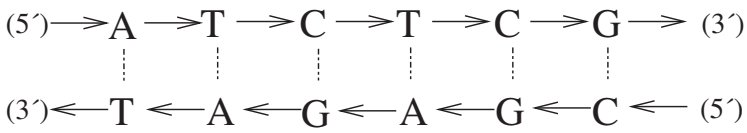

Figura 2.5: Figura esquemática de duas fitas de DNA — antiparalelas e complementares.

As duas fitas carregam portanto a mesma informação. Esta redundância forma o DNA muito mais robusto que o RNA, e permite o reparo automático de regiões danificadas de uma fita.

Em virtude deste modelo, dizemos que dois trechos de DNA são *complementares* caso sejam formados por bases complementares na ordem inversa quando lidos da ponta $5'$ para $3'$.

2.4 Cromossomos

Numa célula de eucarioto, o DNA encontra-se normalmente na forma de algumas dezenas de estruturas compactas chamadas de *cromossomos*. Esta organização provavelmente evoluiu nos eucariotos, cujo DNA é muito extenso, para reduzir o risco de enozamento e quebra. Já nos procariotos, o DNA de uma célula geralmente consiste de uma única fita dupla circular (com as extremidades unidas). Por extensão, esta fita recebe o nome de cromossomo, apesar de não ter a estrutura compacta dos cromossomos de eucariotos. Muitos procariotos possuem cromossomos secundários, chamados *plasmídeos*, com a mesma estrutura (fita dupla circular) do cromossomo principal, porém com tamanho menor.

O conjunto completo dos cromossomos de uma espécie é chamado de *genoma* [14]. O tamanho do genoma varia muito entre as diversas espécies biológicas. O genoma humano, por exemplo, possui aproximadamente 3 bilhões de bases, organizadas em 23 pares de cromossomos [10]. O cachorro possui 39 pares, enquanto a mosca da fruta (*Drosophila melanogaster*) possui apenas 4 pares [6].

O DNA dos procariotos é composto em sua maioria somente por genes. Já os genomas eucariotos contêm uma pequena proporção de genes, sendo o restante composto por regiões intergênicas, que não codificam proteínas.

2.5 Síntese protéica

A informação contida no DNA é utilizada na síntese de proteínas, em um processo chamado *tradução*. Antes disso, porém, ela precisa ser *transcrita* para RNA. Alguns vírus (retrovírus) utilizam a transcrição, neste caso chamada de *transcrição reversa* (de RNA para DNA), no sentido contrário para invadir o genoma da célula hospedeira.

Junto com a *replicação*, que é a capacidade do DNA de formar novas moléculas de DNA, a tradução, a transcrição e a transcrição reversa definem o fluxo da informação genética na célula [14]. Este modelo representado esquematicamente na figura 2.6 é chamado de *dogma central da biologia molecular*.

2.5.1 Transcrição

Na transcrição, uma nova fita de RNA é fabricada a partir de um trecho de uma fita de DNA. Este processo começa quando uma enzima chamada *RNA polimerase*, que existe livre na célula,

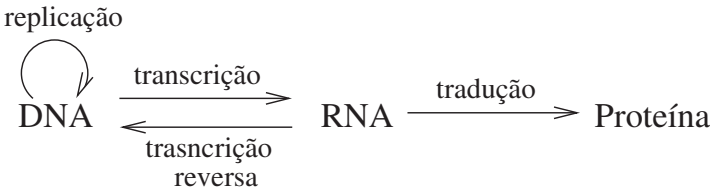

Figura 2.6: Esquema do dogma central da biologia molecular.

se liga a uma das fitas do DNA em uma região específica chamada de *promotor*. A RNA polimerase desfaz as pontes de hidrogênio que unem as duas fitas de modo a expor seus nucleotídeos, e constroi a fita de RNA juntando um nucleotídeo de cada vez. A polimerase escolhe os nucleotídeos respeitando o pareamento de bases complementares, de modo que a seqüência de nucleotídeos do RNA é complementar ao trecho de DNA processado — exceto pelo uso de uracila (U) no lugar da timina (T). Especificamente, cada segmento de uma molécula de DNA que é transcrito para uma molécula de mRNA funcional constitui um *gene* e contém a informação para produzir proteínas.

Nos eucariotos, a transcrição do DNA para mRNA é realizada no núcleo da célula, sendo que a síntese da proteína ocorre somente depois que o mRNA passa ao citoplasma. Porém, antes de deixar o núcleo, o mRNA — que recebe o nome de *transcrito primário* — sofre algumas modificações, transformando-se no chamado *transcrito maduro*. Essas modificações incluem um processo de recorte e emenda da cadeia de mRNA chamado de *splicing*.

Existem portanto segmentos do mRNA primário, os *íntrons*, que são removidos do mRNA pelo *splicing*, e não são usados na tradução de proteínas. Os segmentos restantes, chamados de *éxons*, são unidos para formar o mRNA maduro. Freqüentemente, há variação no processo de *splicing*, de modo que o mesmo mRNA primário pode ser processado (pela combinação de diferentes éxons) de várias formas diferentes, dando origem a diversos mRNA maduros. Esse fenomêno, muito comum nos eucariotos, é chamado de *splicing alternativo*. Portanto, é necessário definir um éxon como sendo um trecho que *pode* ser transcrito no mRNA maduro; e um íntron como um trecho que é *sempre* eliminado.

2.5.2 Tradução

Na etapa de tradução, as proteínas são fabricadas a partir da informação contida no mRNA.

2.5.2.1 Códons

O mRNA é processado pelos componentes celulares em grupos de três nucleotídeos consecutivos, chamados de *códons*. Cada códon especifica um aminoácido, componente fundamental das proteínas.

Neste processo, há a participação do tRNA, cuja molécula possui um sítio em sua ponta 3' para a ligação de um aminoácido; e perto da ponta oposta, uma seqüência de três nucleotídeos chamada de *anticódon* que determinam o aminoácido carregado pelo tRNA. Estes nucleotídeos do anticódon pream-se através da complementariedade de bases a um códon específico do mRNA.

2.5.2.2 Ribossomos

A montagem da proteína é feita por uma estrutura celular chamada *ribossomo*. O ribossomo consiste de um conjunto de proteínas ligadas por um "esqueleto" de RNA (rRNA). Para traduzir uma fita de mRNA, o ribossomo liga-se à mesma num ponto específico, marcado por uma seqüência determinada de bases — geralmente o códon AUG (chamado de *códon de início*). Durante a síntese, o ribossomo desloca-se ao longo da fita de mRNA códon por códon. A medida que os códons são processados, o respectivo tRNA é atraído e ligado ao mRNA. O ribossomo então corta o aminoácido carregado pelo tRNA, e concatena-o à cadeia crescente da proteína, liberando o resíduo do tRNA (que é reciclado por outras enzimas).

A tabela 2.2 exibe os aminoácidos correspondentes a cada códon. Tal correspondência é chamada de *código genético*.

2.5.2.3 Códons de parada

Observa-se que os códons UAA, UAG e UGA não são utilizados na fabricação de aminoácidos. Quando o ribossomo atinge um desses códons, ele se desprende da fita de mRNA e da cadeia de proteína sintetizada. Por essa razão, estas combinações são chamadas de *códons de parada*.

2.5.2.4 Fase de leitura

Os nucleotídeos de uma mesma seqüência de RNA podem ser agrupados em códons de três modos distintos, chamadas de *fases ou quadros de leitura*. Um exemplo pode ser visto na figura 2.7. Na primeira fase de leitura, os códons são combinados a partir do primeiro nucleotídeo da seqüência, enquanto na segunda fase e na terceira, a partir do segundo e do terceiro nu-

Primeira base do códon	Segunda base do códon				Terceira base do códon
	U	C	A	G	
U	Fenilalanina	Serina	Tirosina	Cisteína	U
	Fenilalanina	Serina	Tirosina	Cisteína	C
	Leucina	Serina	(Parada)	(Parada)	A
	Leucina	Serina	(Parada)	Triptofano	G
C	Leucina	Prolina	Histidina	Arginina	U
	Leucina	Prolina	Histidina	Arginina	C
	Leucina	Prolina	Glutamina	Arginina	A
	Leucina	Prolina	Glutamina	Arginina	G
A	Isoleucina	Treonina	Asparagina	Serina	U
	Isoleucina	Treonina	Asparagina	Serina	C
	Isoleucina	Treonina	Lisina	Arginina	A
	Metionina	Treonina	Lisina	Arginina	G
G	Valina	Alanina	Ácido Aspártico	Glicina	U
	Valina	Alanina	Ácido Aspártico	Glicina	C
	Valina	Alanina	Ácido Glutâmico	Glicina	A
	Valina	Alanina	Ácido Glutâmico	Glicina	G

Tabela 2.2: Códons e seus respectivos aminoácidos produzidos.

cleotídeo, respectivamente. Começando em qualquer outra base são gerados subconjuntos de códons iguais aos pertencentes a uma destas três fases. Em quase todos os casos somente uma destas formas de leitura produz uma proteína funcional [4]. Uma fase de leitura com o início no códon de início e que não possua o códon de fim é chamado de *quadro de leitura aberta* ou ORF.

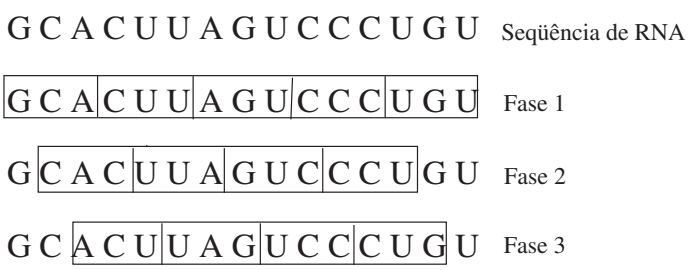

Figura 2.7: Três possíveis fases de leitura para um trecho de RNA.

Capítulo 3

Revisão de probabilidade e estatística

Este capítulo tem por finalidade apresentar conceitos básicos de probabilidade; em particular, o teorema de Bayes, que é utilizado na técnica proposta neste trabalho (e em praticamente todas as aplicações que exigem dedução estatística). Desta forma, na seção 3.1 apresentamos os conceitos de independência e exclusão; na seção 3.2, a definição de probabilidade condicional e na seção 3.3, o teorema de Bayes.

3.1 Independência e exclusão

Considera-se um experimento aleatório que pode ter vários resultados distintos (*eventos*). Seja A um desses eventos: denotaremos por $\Pr(A)$ a probabilidade desse evento ocorrer. Se A e B são dois eventos que podem a princípio ocorrer neste experimento, denotaremos por $\Pr(A \wedge B)$ a probabilidade deles ocorrerem juntos e $\Pr(A \vee B)$ a probabilidade de A ou B (ou ambos) ocorrer. Denotaremos também por $\Pr(\neg A)$ a probabilidade de A não ocorrer, ou seja $1 - \Pr(A)$ [16].

Por definição, dois eventos A e B são independentes se e somente se $\Pr(A \wedge B) = \Pr(A)\Pr(B)$; e são *mutuamente exclusivos* se $\Pr(A \wedge B) = 0$. Se $A_1, A_2 \ldots A_n$ são eventos possíveis de um experimento tais que $\Pr(A_1 \vee A_2 \vee \cdots \vee A_n) = 1$, dizemos que eles são *exaustivos*. Se eles também forem exclusivos dois a dois, temos que $\Pr(A_1) + \Pr(A_2) + \cdots + \Pr(A_n) = 1$.

3.2 Probabilidade condicional

A *probabilidade condicional de A dado B*, denotada por $P(A|B)$, representa a probabilidade do evento A ocorrer, nos experimentos em que o evento B também ocorre. A probabilidade

condicional pode ser obtida por:

$$\Pr(A|B) = \frac{\Pr(A \wedge B)}{\Pr(B)} \tag{3.1}$$

Note que $\Pr(A|B)$ é indefinida se $\Pr(B) = 0$.

3.3 Teorema de Bayes

Seja A_1, A_2, ..., A_k uma coleção de eventos mutuamente exclusivos e exaustivos de um experimento S, e B um evento qualquer associado a S, tal que $\Pr(B) > 0$. Verifica-se a partir da Equação 3.1 que:

$$\Pr(A_i|B) = \frac{\Pr(B|A_i)\Pr(A_i)}{\Pr(B)} \qquad \text{para} \quad i = 1, 2, ..., k \ . \tag{3.2}$$

Por outro lado, verifica-se também que:

$$\Pr(B) = \sum_{j=1}^{k} \Pr(B|A_j)\Pr(A_j)$$

Portanto, a fórmula 3.2 pode ser escrita como:

$$\Pr(A_i|B) = \frac{\Pr(B|A_i)\Pr(A_i)}{\sum_{j=1}^{k} \Pr(B|A_j)\Pr(A_j)} \qquad \text{para} \quad i = 1, 2, ..., k \ . \tag{3.3}$$

Os termos $\Pr(A_i)$ e $\Pr(A_i|B)$ são chamados respectivamente, probabilidades *a priori* e *a posteriori* para o evento A_i.

A fórmula de Bayes é muito utilizada em várias áreas com aplicações estatísticas tais como medicina e biologia. Tipicamente, nesses casos os eventos A_1, A_2, ..., A_k representam "causas" possíveis e B um "efeito" observado. Conhecem-se as probabilidades $\Pr(B|A_i)$ de cada causa A_i produzir o efeito B; e deseja-se calcular $\Pr(A_i|B)$ — a probabilidade de cada causa A_i ter ocorrido, dado que o efeito B foi observado.

Capítulo 4

O problema da identificação de regiões codificadoras

Neste capítulo, decrevemos mais detalhadamente o problema da identificação de regiões codificadoras, e fazemos uma revisão dos principais trabalhos relacionados existentes na literatura. Na seção 4.1 é dada a definição usual para o problema e na seção 4.2 são apresentados alguns problemas da identificação experimental que motivam a identificação computacional (apresentada em 4.3). Na seção 4.4 é apresentada a definição para o problema utilizada neste trabalho e na seção 4.5 a representação adotada para sua solução. Finalmente na seção 4.6, fazemos uma revisão bibliográfica dos principais trabalhos existentes na literatura.

4.1 Definição usual para o problema

O primeiro passo na interpretação de um genoma é a identificação de genes — trechos que são transcritos para mRNA — e de regiões codificadoras — trechos que são traduzidos para proteínas. Esta identificação é relativamente simples nos procariotos, onde cada gene é uma única região codificadora e os genes cobrem praticamente todo o genoma. Neste caso, o problema é apenas localizar o início e o final de cada gene. Já nos eucariotos essa tarefa é mais complexa, uma vez que dentro de um gene pode existir uma grande quantidade de regiões codificadoras (éxons), interrompidas por regiões não codificadoras (íntrons).

4.2 Identificação experimental

A identificaçao definitiva de regiões codificadoras é feita experimentalmente, isolando-se as moléculas de mRNA presentes nas células e determinando sua seqüência de DNA. Comparando-

se a seqüência do mRNA com o DNA é possível detectar os genes, e seus éxons e íntrons.

Este método é muito trabalhoso e demorado. Além disso, certos genes (ou éxons) não são transcritos o tempo todo; nesses casos, é necessário adivinhar as circunstâncias em que eles são ativados, o que nem sempre é possível. Outra dificuldade é a existência de *pseudo-genes* nos cromossomos — trechos que possuem a mesma estrutura de um gene normal, mas não produzem uma proteína funcional, geralmente devido a alguma leve mutação. Além disso, há o *splicing* alternativo nos eucariotos, no qual a partir de um mesmo trecho de DNA podem ser produzidos dezenas de mRNAs distintos.

4.3 Identificação computacional

Devido a todas essas dificuldades do método experimental, há grande interesse em métodos puramente computacionais, que identificam genes e regiões codificadoras usando apenas a seqüência de bases do genoma (DNA, ou, no caso de retrovírus, RNA).

Este problema vem sendo estudado há pelo menos vinte anos. Dentre os diversos métodos computacionais descritos para tal fim, as duas principais estratégias são *homologia* e *ab initio*.

Nos métodos baseados em homologia, os éxons, ou os genes e sua estrutura, são deduzidos pela sua semelhança com seqüências de aminoácidos, nucleotídeos ou proteínas em genomas previamente analisados e classificados. Obviamente esta abordagem não permite encontrar "novos" genes, ou genes que diferem substancialmente de seus homólogos em outras espécies.

Os métodos *ab initio* usam dois tipos de informação: *sinais* e *conteúdo*. Os sinais são seqüências e padrões específicos que indicam o início ou fim de trechos codificadores, como promotores, códons de parada, marcadores de *splicing* etc. A informação de conteúdo explora o fato que as regiões codificadoras, ao contrário das não-codificadoras, possuem função biológica, o que induz a certos padrões e periodicidades nas seqüências.

4.4 Definição utilizada para o problema

Nesta dissertação estudamos um método computacional *ab initio* para detecção de regiões codificadoras, baseado em conteúdo — especificamente na distribuição estatística das bases (e de grupos de bases consecutivas).

Esta abordagem é inerentemente incapaz de distinguir genes de pseudo-genes — ou, mais geralmente, regiões codificadoras ativas (que ainda são utilizadas pelo organismo) de regiões

codificadoras fosséis — que não codificam proteínas, mas que já foram um dia ativas. Portanto, temos que mudar a definição do problema para a detecção de regiões codificadoras — entendendo-se que estas últimas podem ser tanto ativas quanto fósseis, sendo que nosso método não é capaz de distinguí-las.

Uma vez que a identificação é feita apenas baseada na distribuição de seqüências de bases, de modo geral, ela nunca pode oferecer certeza absoluta na classificação. Como veremos, neste modelo praticamente todo trecho de DNA pode ser interpretado como sendo codificador ou não-codificador — apenas as probabilidades são diferentes.

4.4.1 Modelo de DNA utilizado

Consideramos o DNA como uma concatenação de trechos (regiões) de dois tipos: regiões codificadoras de proteínas — formadas pela junção de códons — e de regiões não-codificadoras. É importante resaltar que no nosso modelo todo éxon é uma região codificadora, mas o contrário não é verdade. Regiões codificadoras (fósseis) podem ocorrer nos íntrons ou nas regiões intergênicas. O modelo de DNA considerado é visto esquematicamente na figura 4.1.

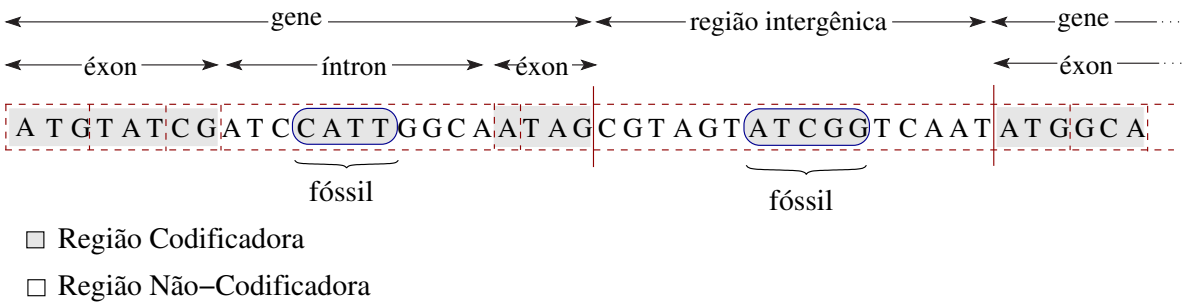

Figura 4.1: Estrutura esquemática do modelo de DNA utilizado.

Para simplificação dos cálculos, supomos, como Staden e McLachlan [22] que os códons nos trechos codificadores são aleatórios e independentes. Porém, observa-se que códons diferentes ocorrem com probabilidades diferentes, o que resulta em uma dependência entre os nucleotídeos que compõem cada códon.

Por outro lado, supomos que as bases individuais nas seqüências não-codificadoras são aleatórias e independentes entre si e que suas probabilidades são diferentes das probabilidades nas regiões codificadoras. Estas diferenças nos permitirão identificar os dois tipos de seqüências. Supomos também que as regiões intergênicas possuem estas mesmas composições estatísticas das regiões não-codificadoras.

Este modelo das regiões não-codificadoras é uma simplificação muito grosseira. Por ex-

emplo, sabe-se que as regiões não-codificadoras contêm freqüentemente trechos com grande quantidade de repetições, bem como regiões codificadoras complementadas, promotores e outras seqüências com função biológica importante — que induzem dependências entre bases vizinhas.

Como veremos, a abordagem descrita acima fornece, na maioria dos casos, também a fase de leitura das regiões identificadas como codificadoras.

4.4.2 Analogia do problema com a identificação de idiomas

Para um melhor entendimento do problema de identificação de regiões codificadoras podemos fazer uma analogia deste com o problema de identificação estatística de idiomas. Suponha que temos um texto que, sabidamente, é uma mistura de trechos em português e trechos em inglês — em maiúsculo, sem espaços ou acentos, como na figura 4.2; e, dado apenas o texto misturado, queremos identificar os trechos do mesmo que pertencem a cada língua.

Inglês	Português	Misto
BATSAREUGLYANIMALS	GOSTODEFRUTAS	GTASQUAISQUERFRU INGI
YESTERDAYEVENINGI	QUAISQUERFRUTASBANANAS	HEARDTHATJOHNSAWABA N
HEARDTHATJOHNSAWABAT	SAODELICIOSASLARANJAS	ANASSAODELICIOSASLAR
INTHECELLAROFHIS	TAMBEMMACASABACAXISE	ANJASTAM THEBATFLEW
HOUSETHEBATFLEWAWAY	MANGAS	AWAYASJOHNENT XISEMAN
ASJOHNENTERED		

Figura 4.2: Trechos de texto em português, em inglês e misturado.

Intuitivamente notamos que esse problema é solúvel desde que os trechos sejam relativamente longos. Se examinarmos uma única letra, como na figura 4.3-A, a classificação será pouco conclusiva, uma vez que os dois idiomas são compostos pelo mesmo alfabeto.

A identificação fica mais conclusiva quando consideramos duas ou mais letras consecutivas como na figura 4.3-B. A letra “A” na seqüência “SAO” provavelmente pertence a um trecho em português, porque “SAO” é muito mais comum na língua portuguesa do que na inglesa — apesar de que cada letra (S,A,O) por si não permite identificar a língua com confiança. Entretanto, a identificação ainda não é categórica, pois “SAO” pode ser parte de frases em inglês como “HIS AORTA”; mas essas frases são bastante raras.

De modo geral, a identificação fica tanto mais clara quanto maior for o trecho analisado, como na figura 4.3-C. Olhando para o trecho “ANASSAODELI”, constatamos que o mesmo contém vários grupos de letras que são mais comuns em português do que em inglês. Portanto, mesmo sem entender o significado, podemos afirmar com relativa segurança que o “A” central

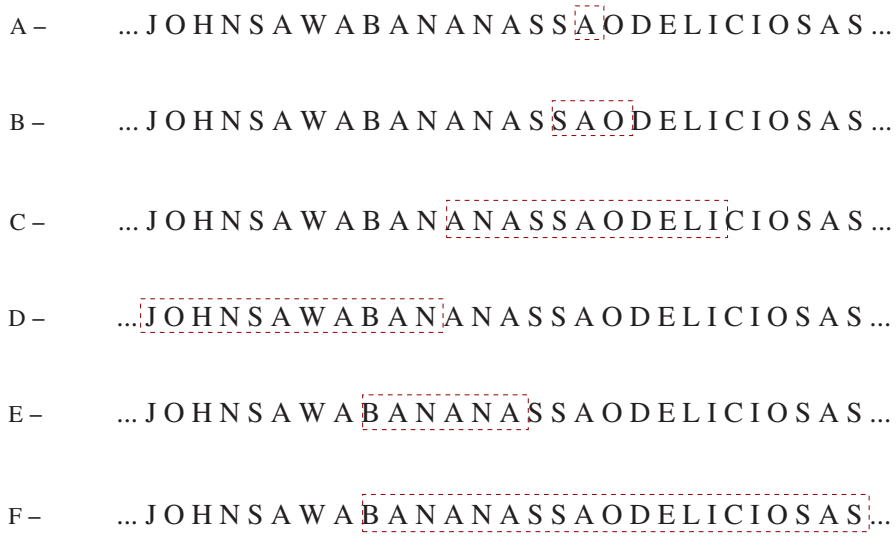

Figura 4.3: Trechos de texto misturado.

pertence a um trecho de texto em português. Por outro lado, no trecho "JOHNSAWABAN" da figura 4.3-D, constatamos vários grupos de letras mais típicos de inglês do que de português. Assim, podemos afirmar com relativa segurança que a letra central "A" desse trecho provém de um texto em inglês.

Podemos notar por estes dois últimos exemplos que o trecho "BANANA" (figura 4.3-E) é inerentemente ambíguo, podendo pertencer tanto a um trecho em português, quanto em inglês. Esta ambiguidade persiste por maior que seja o tamanho da janela considerada. Este exemplo mostra que em geral é impossível determinar com precisão onde está a fronteira entre trechos de línguas diferentes, e ter certeza absoluta sobre a identificação de um trecho. Como veremos, a identificação *ab initio* de regiões codificadoras do DNA tem a mesma limitação inerente.

Neste exemplo, a identificação dos trechos em português é reforçada pelo fato das sentenças "JOHNSAW" e "SAODELICIOSAS" terem sentido em inglês e português respectivamente. Usar esta informação equivale, no caso do DNA, a usar homologia com seqüências de bases ou função de proteína conhecidas pelo trecho de DNA em questão.

Note que o significado da frase "BANANASSAODELICIOSAS" da figura 4.3-F só é compreendido depois que separarmos as palavras "BANANAS", "SAO" e "DELICIOSAS". Por outro lado, as mesmas letras podem ser divididas em palavras de outras formas, com sentido ou não: "ABA NA", "ASSA", "CIOSA" etc. Entretanto, note que é possível identificar com confiança a língua como português, mesmo sem ter certeza da divisão das palavras, apenas pelas freqüências de letras e grupos de letras.

Do mesmo modo, em alguns trechos de DNA é possível dizer com confiança que uma

região é codificadora, mesmo quando não há muita certeza sobre a fase de leitura (divisão em códons). Porém, ao contrário da maioria das línguas naturais, no DNA estes casos são raros. Em geral, quando nosso método identifica uma região codificadora com boa segurança, ele também identifica a fase de leitura da mesma, também com boa segurança.

4.5 Representação da solução

Idealmente, uma ferramenta de classificação deve rotular uma seqüência de DNA fornecida com os rótulos “C” (codificadora) ou “N” (não-codificadora). Porém, em um método baseado na análise estatística da seqüência, os resultados são obtidos inicialmente na forma de uma probabilidade para cada rótulo. Se insistirmos em escolher apenas um dos rótulos para colocar na saída, estaremos jogando fora informação — o valor da probabilidade — que pode ser útil para o usuário, pois revela a confiança que o programa tem na sua classificação. Portanto, é melhor fornecer como resultado as probabilidades dos dois rótulos, em vez de dar apenas o rótulo mais provável.

4.6 Revisão bibliográfica

4.6.1 Identificação por homologia

Métodos baseados em homologia [5, 8], são eficazes em certas aplicações, como o reconhecimento de pequenos éxons com freqüências de ocorrência incomuns. Entretanto, eles não servem para encontrar éxons pertencentes a novos genes, sem seqüências homólogas conhecidas. Sendo também muito sensíveis a erros de fase, e exigem seqüências homólogas de organismos não muito distantes na escala evolucionária.

Para tentar resolver estes problemas, alguns autores combinam a análise por homologia com outras técnicas. Assim, Rogozin, D’Angelo e Milanesi [18] tentam utilizar outras informações — a análise de propriedades evolucionárias conservadas das seqüências codificadoras — quando a seqüência procurada não contém regiões homólogas conhecidas. Segundo os autores o método possui uma boa taxa de reconhecimento, apesar de ainda apresentar alguns problemas como a sensibilidade a certos tipos de mutação. Não vamos detalhar mais as abordagens que utilizam puramente homologia pois não é o tema deste trabalho.

4.6.2 Identificação *ab initio*

Fickett, um dos precursores no trabalho do reconhecimento das regiões codificadoras, descreve em 1982 [7] um algoritmo chamado TESTCODE que utiliza a variação da freqüência das bases com sua posição dentro do códon para tentar identificar regiões codificadoras e não-codificadoras. Treinando o TESTCODE com metade do banco de dados de seqüências disponíveis e classificando a outra metade, o autor verificou que o algoritmo erra em apenas 5% das seqüências testadas e não consegue retornar opinião alguma em 18% das seqüências — ou seja, conseguiu uma taxa de acerto de 77%. Porém, o TESTCODE não funciona bem para seqüências menores que 200 bases e, segundo o autor, é completamente insensível na detecção da fase. Uma limitação mais séria do método é que ele não prevê a presença de trechos codificadores nas mesmas seqüências: ele apenas tenta classificar a seqüência inteira como codificadora ou não. Outros trabalhos pioneiros no uso de técnicas estatísticas foram o de R.Staden e A.D.McLachlan [22], e o trabalho de M.Gribskov, J.Devereux e R.R.Burgess [9].

Já em 2002, para tentar reconhecer seqüências codificadoras com tamanhos pequenos, especialmente as pertencentes a genes humanos, Y.Wang, C-T.Zhang e P.Dong [23] utilizam quatro medidas. Em uma delas, os autores medem a assimetria na distribuição das bases dentro das três posições do códon, em outra, uma peridiocidade de 3 bases presente nas seqüências codificadoras. Esta última característica é detectada tratando-se a seqüência como um sinal e aplicando-se a transformada de Fourier, adaptada especialmente para capturar tais propriedades em seqüências curtas. Os autores utilizam ainda a informação de que a base predominante na primeira posição do códon é uma purina, medindo sua freqüência de ocorrência nas três fases de leitura. E por fim, é feito um estudo da ocorrência dos códons de parada (TAG, TAA e TGA). Após a introdução das quatro características, cada seqüência é representada por um ponto ou um vetor no espaço de dimensão 4. Para distinguir entre seqüências codificadoras e não-codificadoras ou codificadoras e regiões inter-gênicas é usado o algoritmo discriminante linear de Fisher. Em testes o método proposto apresentou bons resultados especialmente se comparado a outros, fato este que os autores atribuem à detecção dos códons de parada. Porém, como sabemos nem sempre tal informação é disponível quando obtemos um trecho arbitrário de uma seqüência de DNA.

Wu *et al.* [24] também procuram reconhecer éxons (e íntrons) de tamanho pequeno em humanos. Para isto, os autores utilizam um conjunto de três características estatísticas, chamadas de SZ, juntamente com o reconhecimento dos códons de parada. As características capturadas pelo método são baseadas nas diferenças individuais das freqüências dos nucleotídeos nas três posições do códon nas ORFs e nas regiões não-codificadoras. Essas características são maxi-

mizadas com o objetivo de capturar a correta fase de leitura. As características SZ podem ser visualizadas em três ou duas dimensões. Para seqüências de tamanho menor que 140 pares de bases, as características SZ obtêm um taxa de reconhecimento acima de 92%.

Recentemente, uma grande quantidade de programas para o reconhecimento das regiões codificadoras utilizam Cadeias de Markov de estados ocultos, sendo esta técnica empregada em diversos trabalhos na literatura [13, 15]. Em particular, no trabalho de Majoros, Pertea e Salzberg [13] em 2005, os autores utilizaram uma generalização de uma cadeia de Markov de estados ocultos para obter padrões conservados entre aminoácidos e nucleotídeos em regiões codificadoras e não-codificadoras de espécies próximas na escala evolucionária. Estas informações são incorporadas tanto no processo de predição de genes quanto de regiões codificadoras. O programa apresentou uma ótima taxa de reconhecimento das regiões codificadoras (cerca de 89%). Porém, os dados utilizados nos testes foram filtrados para eliminar seqüências com desigual quantidade de éxons e obter seqüências com similaridade entre aminoácidos maiores do que 90%.

Existem ainda várias outras técnicas empregadas que não serão detalhadas aqui como: transformada de Fourier ([12]), redes neurais [20, 21] e árvore de decisão [19].

Como vimos, existem vários métodos computacionais com bons resultados no reconhecimento das regiões codificadoras, porém quase todos são pouco genéricos funcionando extremamente bem para tipos específicos de éxons, tais como regiões codificadoras de tamanho grande, e não reconhecendo outros tipos, como éxons de tamanho pequeno, ou o contrário. O que continua motivando a busca por novas técnicas cada vez mais precisas.

Capítulo 5

Método proposto

Neste capítulo apresentamos o método proposto para a identificação de regiões codificadoras que, como exposto no capítulo 4, é baseado nas características estatísticas de seqüências de bases consecutivas. Na seção 5.1 fazemos algumas considerações iniciais sobre a terminologia empregada e na seção 5.2 apresentamos o método proposto juntamente com as hipóteses consideradas.

5.1 Considerações iniciais

Estabelecemos algumas notações para facilitar a formulação matemática do problema. Consideramos uma cadeia de DNA de tamanho n como uma seqüência $b = b_1, b_2, \ldots, b_n$ de letras do alfabeto $\Gamma = \{A, C, T, G\}$. Por exemplo, se $b =$ ATGGATATCGGTCGA então $n = 15$ e, $b_1 =$ A, $b_2 =$ T, $\cdots$, $b_{15} =$ A. Segundo o modelo da seção 4.4.1, supomos que cada nucleotídeo b_i em uma cadeia de DNA tem também uma *classificação* ou *rótulo* e_i que é um dos elementos do alfabeto $\Upsilon = \{N, D, E, F\}$. Especificamente, se o nucleotídeo faz parte de uma região não-codificadora, ele possui rótulo N. Se o nucleotídeo faz parte de uma região codificadora, ele tem rótulo D, E ou F, dependendo da sua posição dentro do códon: a primeira base do códon é classificada como D, a base do meio como E, e a base final como F.

Por exemplo, suponha que as seis primeiras bases da cadeia $b =$ ATGGATATCGGTCGA pertencem a uma região codificadora, formada por dois códons inteiros (ATGGAT), e que o restante é uma região não-codificadora. A classificação da seqüência b então é a seguinte:

$$b = \mathsf{ATGGATATCGGTCGA}$$
$$e = \mathsf{DEFDEFNNNNNNNNN}$$

Denotamos por Γ^k e Υ^k, o conjunto de todas as possíveis seqüências de k elementos de Γ e Υ respectivamente. Por exemplo, Γ^3 é um conjunto com $4^3 = 64$ elementos: $\{$AAA, AAT, $\cdots$, GGG$\}$.

Chamamos os elementos de Γ^k e Υ^k de *k-tuplas* e *k-eventos*, respectivamente. Por exemplo, nas seqüências:

$$b = \texttt{ATG}\boxed{\texttt{GATATC}}\texttt{GGTCGA}$$
$$e = \texttt{DEF}\boxed{\texttt{DEFNNN}}\texttt{NNNNNN}$$

A seqüência GATATC é uma 6-tupla de b, e a seqüência DEFNNN é um 6-evento de e.

5.2 Método proposto

Com esta notação, podemos reformular nosso problema da seguinte forma: dada uma seqüência de bases (DNA ou RNA) de tamanho arbitrário, desejamos calcular uma classificação probabilística (conforme a seção 4.5) usando os rótulos $\{N, D, E, F\}$.

Mais precisamente, dada uma seqüência $b = b_1, \ldots, b_n$ calculamos $4n$ probabilidades $\Pr^*[i, \epsilon] = \Pr(e_i = \epsilon)$ para cada $i \in [1, n]$ e cada $\epsilon \in \Upsilon$, onde e_i é o rótulo correto (desconhecido) de b_i.

No método proposto, as probabilidades $\Pr^*$ são calculadas através do Teorema de Bayes, a partir das probabilidades de bases e de códons com rótulos dados, usando algumas hipóteses simplificadas. Mais precisamente, calculamos as probabilidades dos vários k-eventos que podem ser associados a cada k-tupla da cadeia dada, onde k é um parâmetro especificado pelo usuário, também chamado de *janela*.

Esta etapa de *classificação* é precedida de uma etapa de *treinamento* na qual as freqüências de bases e códons são obtidas analisando amostras de DNA com rotulação conhecida.

Idealmente, a classificação deveria ser *conclusiva*; isto é, $\Pr^*[i, e_i] \approx 1$, e $\Pr^*[i, \epsilon'] \approx 0$ para todo $\epsilon' \neq e_i$. Entretanto, como as seqüências codificadoras e não-codificadoras usam o mesmo alfabeto, diferindo apenas nas freqüências das letras (e dos grupos de letras), essa resposta ideal geralmente é impossível de se obter.

5.2.1 Classificação baseada em contexto

Como vimos na seção 4.4.2, para uma classificação confiável de cada nucleotídeo $b_i \in b$ é necessário um exame de contexto, ou seja, a análise de seus nucleotídeos vizinhos. Portanto,

para classificar cada base b_i da seqüência examinamos t bases imediatamente antes de b_i e t bases imediatamente depois. Ou seja, examinamos um número ímpar $k = 2t + 1$ de bases consecutivas, para classificar a base central. Para cada nucleotídeo b_i, com $i = \{t, \ldots, n - t\}$, calculamos as probabilidades $\Pr^*[i, \epsilon]$ da seguinte forma:

$$\Pr^*[i, \epsilon] = \sum_{\beta'\beta''} \Pr(\beta = \beta'\epsilon\beta''|\alpha_i) \tag{5.1}$$

onde α_i é a k-tupla centrada em b_i, β_i é seu k-evento desconhecido e β' e β'' variam sobre todos os t-eventos possíveis.

5.2.2 Uso da fórmula de Bayes

Cada termo do somatório da equação 5.1 é a probabilidade da k-tupla α_i da seqüência b ter a classificação $\beta = \beta'\epsilon\beta''$. Calculamos essa probabilidade pela fórmula de Bayes (dada na equação 3.2), da seguinte forma:

$$\Pr(\beta|\alpha_i) = \frac{\Pr(\alpha_i|\beta)\Pr(\beta)}{\Pr(\alpha_i)} \tag{5.2}$$

Nesta fórmula, $\Pr(\alpha_i|\beta)$ é a probabilidade de uma k-tupla de DNA com classificação β possuir a seqüência de bases α_i; $\Pr(\beta)$ é a probabilidade *a priori* de uma k-tupla ter a classificação β; e $\Pr(\alpha_i)$ é a probabilidade de uma k-tupla de DNA ser constituida pela seqüência de bases α_i, calculada pelo somatório de todos os $\Pr(\alpha_i|\beta)\Pr(\beta)$.

5.2.3 Hipóteses sobre estatísticas do DNA

No método proposto, supomos que as seqüências codificadoras e não-codificadoras possuem certas diferenças nas suas propriedades estatísticas, apresentadas nas próximas seções, que permitem a sua identificação. Em todas as análises estatísticas, usamos o conjunto de seqüências de teste de Martin Reese [1] descrito na seção 6.3.

5.2.3.1 Regiões codificadoras

Segundo nosso modelo de DNA, nas regiões codificadoras os nucleotídeos não ocorrem com igual probabilidade, e nem independentemente de seus nucleotídeos vizinhos.

Estas características das regiões codificadoras podem ser observadas nas tabelas 5.1 e 5.2. A coluna “Geral” da tabela 5.1 fornece as freqüências de cada base (A, T, C e G) observadas

nas regiões codificadoras ativas (éxons). As outras três colunas mostram as probabilidades de cada base nas três posições dentro dos códons (D, E e F).

	Geral	**D**	**E**	**F**
A	0,231	0,248	0,300	0,144
T	0,203	0,160	0,263	0,184
C	0,283	0,261	0,235	0,351
G	0,283	0,330	0,200	0,318

Tabela 5.1: Probabilidades de ocorrência dos nucleotídeos A, T, C e G nas regiões codificadoras do DNA, em qualquer uma das posições (coluna Geral) e em cada posição dentro do códon (colunas D, E, F).

A explicação para a tabela 5.1 é que os 64 códons possíveis não ocorrem com igual probabilidade, o que leva a relações de dependência entre os nucleotídeos do mesmo códon. Este fato pode ser observado na tabela 5.2. A coluna $\Pr$ dessa tabela mostra as freqüências relativas de cada códon nas regiões codificadoras. Estes valores podem ser comparados à coluna $\Pr_i$, que mostra as probabilidades de cada códon, estimada supondo que as três bases são escolhidas independentemente e sem levar em conta a posição, calculadas com as freqüências da coluna "Geral" da tabela 5.1. A coluna $\Pr_d$ exibe outra estimativa, calculada da mesma forma mas levando em consideração a posição da base no códon (colunas D, E e F da tabela 5.1) e considerando a independência dos nucleotídeos vizinhos.

Como podemos observar, os valores de $\Pr_i$ e de $\Pr_d$ não são muito semelhantes aos de $\Pr$, com isto, podemos considerar que os nucleotídeos no códon são dependentes de seus vizinhos.

Por outro lado, supomos que os códons são independentes uns dos outros, como proposto por Staden e Mclachlan [22].

5.2.3.2 Regiões não-codificadoras

A tabela 5.3 mostra as freqüências relativas observadas nas regiões não codificadoras da base de dados.

Para as regiões não-codificadoras, supomos que cada base foi gerada independentemente das demais com a probabilidade independente da posição, exibida na tabela 5.3.

Como observado na seção 4.4.1, este modelo é uma simplificação grosseira sem justificativa experimental adequada, mas adotado por questões de simplicidade. Infelizmente não temos condições de melhorar este modelo, principalmente porque, nas bases de dados que temos, as regiões codificadoras fosséis não estão identificadas — temos apenas a classificação em éxons e íntrons, sendo que estes últimos podem conter seqüências codificadoras fosséis.

Códon	Pr	Pr_i	Pr_d
AAA	0,018	0,012	0,011
AAT	0,013	0,011	0,011
AAC	0,021	0,015	0,026
AAG	0,035	0,015	0,024
ATA	0,004	0,011	0,009
ATT	0,012	0,010	0,012
ATC	0,024	0,013	0,023
ATG	0,022	0,013	0,021
ACA	0,011	0,015	0,009
ACT	0,011	0,013	0,011
ACC	0,023	0,018	0,021
ACG	0,007	0,019	0,019
AGA	0,008	0,015	0,007
AGT	0,008	0,013	0,009
AGC	0,020	0,019	0,018
AGG	0,010	0,019	0,016
TAA	0,001	0,011	0,007
TAT	0,010	0,010	0,009
TAC	0,018	0,013	0,017
TAG	0,001	0,013	0,015
TTA	0,003	0,010	0,006
TTT	0,013	0,008	0,008
TTC	0,024	0,012	0,015
TTG	0,010	0,012	0,006
TCA	0,008	0,013	0,005
TCT	0,011	0,012	0,007
TCC	0,018	0,016	0,013
TCG	0,005	0,016	0,012
TGA	0,002	0,013	0,005
TGT	0,008	0,012	0,006
TGC	0,014	0,016	0,011
TGG	0,014	0,016	0,010

Códon	Pr	Pr_i	Pr_d
CAA	0,010	0,015	0,011
CAT	0,008	0,013	0,014
CAC	0,015	0,018	0,028
CAG	0,036	0,019	0,025
CTA	0,005	0,013	0,010
CTT	0,010	0,012	0,013
CTC	0,022	0,016	0,024
CTG	0,051	0,016	0,022
CCA	0,014	0,018	0,009
CCT	0,018	0,016	0,011
CCC	0,025	0,023	0,022
CCG	0,008	0,023	0,020
CGA	0,006	0,019	0,008
CGT	0,005	0,016	0,010
CGC	0,016	0,023	0,018
CGG	0,013	0,023	0,017
GAA	0,023	0,015	0,014
GAT	0,018	0,013	0,018
GAC	0,028	0,019	0,035
GAG	0,045	0,019	0,032
GTA	0,005	0,013	0,013
GTT	0,008	0,012	0,016
GTC	0,016	0,016	0,031
GTG	0,033	0,016	0,028
GCA	0,013	0,019	0,011
GCT	0,019	0,016	0,014
GCC	0,036	0,023	0,027
GCG	0,010	0,023	0,025
GGA	0,014	0,019	0,010
GGT	0,012	0,016	0,012
GGC	0,030	0,023	0,023
GGG	0,017	0,023	0,021

Tabela 5.2: Probabilidades de ocorrência dos códons nas regiões codificadoras do DNA.

A tabela 5.4 mostra o problema. A coluna Pr_o exibe a freqüência observada de cada 3-tupla de nucleotídeos extraídas nos íntrons da base de dados. A coluna Pr_i fornece as probabilidades calculadas pelo nosso modelo. As diferenças entre as duas colunas revelam que existe dependência entre bases vizinhas nas regiões não-codificadoras.

5.2.4 Aproximação para trechos grandes

Uma dificuldade na nossa abordagem é que o número de k-eventos β a considerar na fórmula de Bayes é enorme (4^k). Para reduzir a análise a um custo aceitável, usamos duas hipóteses adicionais:

1. Nas cadeias de DNA, entre quaisquer dois trechos C há sempre um trecho N;

Base	$\mathbf{Pr}$
A	0,263
T	0,278
C	0,227
G	0,232

Tabela 5.3: Probabilidades de ocorrência dos nucleotídeos nas regiões não-codificadoras do DNA.

Triplas	$\mathbf{Pr}_o$	$\mathbf{Pr}_i$
AAA	0,030	0,018
AAT	0,019	0,019
AAC	0,012	0,016
AAG	0,018	0,016
ATA	0,015	0,019
ATT	0,020	0,020
ATC	0,012	0,017
ATG	0,016	0,017
ACA	0,017	0,016
ACT	0,015	0,017
ACC	0,013	0,014
ACG	0,003	0,014
AGA	0,020	0,016
AGT	0,016	0,017
AGC	0,016	0,014
AGG	0,021	0,014
TAA	0,017	0,019
TAT	0,015	0,020
TAC	0,010	0,017
TAG	0,012	0,017
TTA	0,017	0,020
TTT	0,034	0,021
TTC	0,019	0,018
TTG	0,018	0,018
TCA	0,018	0,017
TCT	0,022	0,018
TCC	0,018	0,014
TCG	0,003	0,015
TGA	0,019	0,017
TGT	0,020	0,018
TGC	0,016	0,015
TGG	0,021	0,015

Triplas	$\mathbf{Pr}_o$	$\mathbf{Pr}_i$
CAA	0,016	0,016
CAT	0,016	0,017
CAC	0,016	0,014
CAG	0,023	0,014
CTA	0,011	0,017
CTT	0,020	0,018
CTC	0,020	0,014
CTG	0,024	0,015
CCA	0,020	0,014
CCT	0,022	0,014
CCC	0,019	0,012
CCG	0,005	0,012
CGA	0,003	0,014
CGT	0,003	0,015
CGC	0,004	0,012
CGG	0,005	0,012
GAA	0,018	0,016
GAT	0,012	0,017
GAC	0,010	0,014
GAG	0,020	0,014
GTA	0,010	0,017
GTT	0,014	0,018
GTC	0,011	0,015
GTG	0,017	0,015
GCA	0,015	0,014
GCT	0,016	0,015
GCC	0,016	0,012
GCG	0,004	0,012
GGA	0,018	0,014
GGT	0,013	0,015
GGC	0,016	0,012
GGG	0,020	0,012

Tabela 5.4: Probabilidade de ocorrência das triplas de bases nas regiões não-codificadoras do DNA.

2. Os trechos C e N são, em média, bem maiores que o tamanho k da janela.

A hipótese (1) implica que os eventos β que contenham pares DD, DF, ED, EE, FF, FE têm probabilidade *a priori* $\Pr(\beta) = 0$. A hipótese (2) implica que eventos β que contenham mais de uma quebra tem probabilidade $\Pr(\beta)$ muito baixa e podem ser desprezados sem grande efeito no resultado final.

Com estas duas hipóteses, o número de eventos β que devem ser considerados é muito menor que 4^k. Temos apenas as seguintes possibilidades:

1. Evento totalmente não-codificador, neste caso, $\beta = N^k$.
2. Evento totalmente codificador, ou seja $\beta = DEFDEF\ldots$, $\beta = EFDEFD\ldots$ ou $\beta = FDEFDE\ldots$. Totalizando 3 possíveis eventos deste tipo.
3. Evento de transição codificador para não-codificador na forma $N^m C^{k-m}$, com $1 < m < k-1$. Caso existam $k-1$ posições para a quebra, e em cada caso a parte C pode começar com D, E ou F; temos $(k-1) * 3$ eventos deste tipo.
4. Evento de transição não-codificador para codificador na forma $C^m N^{k-m}$ com m como no item anterior; temos $(k-1) * 3$ eventos deste tipo.

Temos portanto apenas $1 + 3 + 3(k-1) + 3(k-1) = 6k - 2$ eventos β a considerar, para cada posição i. Os eventos possíveis para $k = 3$ e $k = 5$ são mostrados nas figuras 5.1 e 5.2, respectivamente.

5.2.5 Decomposição de k-tuplas e k-eventos

Em princípio, as probabilidades $\Pr(\alpha|\beta)$, $\Pr(\beta)$, e $\Pr(\alpha)$ (chamando α_i de α) da equação 5.2 (Bayes) poderiam ser obtidas pela análise de um banco de seqüências rotuladas. Entretanto, verifica-se que para uma classificação confiável k deve ser relativamente grande (dezenas de bases). Neste caso, como o número de k-tuplas α e k-eventos β é muito grande, esta abordagem torna-se impraticável: as tabelas de probabilidades seriam grandes demais, e não haveriam amostras rotuladas suficientes para preencher todas as entradas com valores significativos.

5.2.5.1 Cálculo de $\Pr(\alpha|\beta)$

Resolvemos este problema usando as hipóteses de independência de códons e bases encontradas na seção 5.2.3. Essas hipóteses implicam que podemos decompor $\Pr(\alpha|\beta)$ no produto de probabilidades de j-tuplas e j-eventos com j pequeno.

Especificamente, a hipótese da independência de códons nos permite quebrar $\Pr(\alpha|\beta)$ nas fronteiras dos códons. Por exemplo, para $\alpha =$ ATATCG e $\beta =$ DEFDEF, podemos pela fórmula calcular $\Pr(\text{ATATCG}|\text{DEFDEF}) = \Pr(\alpha_1 = \text{ATA}|\beta_1 = \text{DEF}) \times \Pr(\alpha_2 = \text{TCG}|\beta_2 = \text{DEF})$.

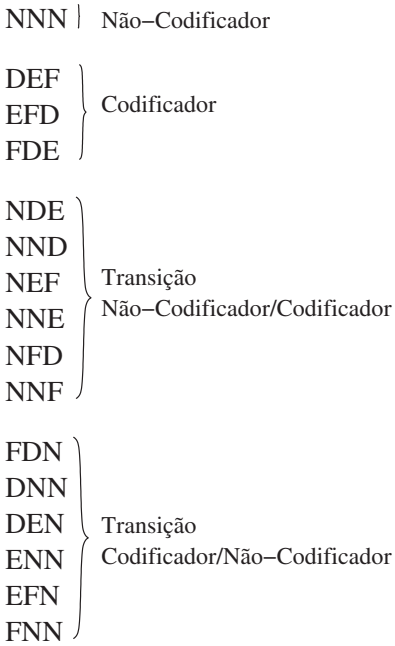

Figura 5.1: 3-eventos considerados para uma 3-tupla.

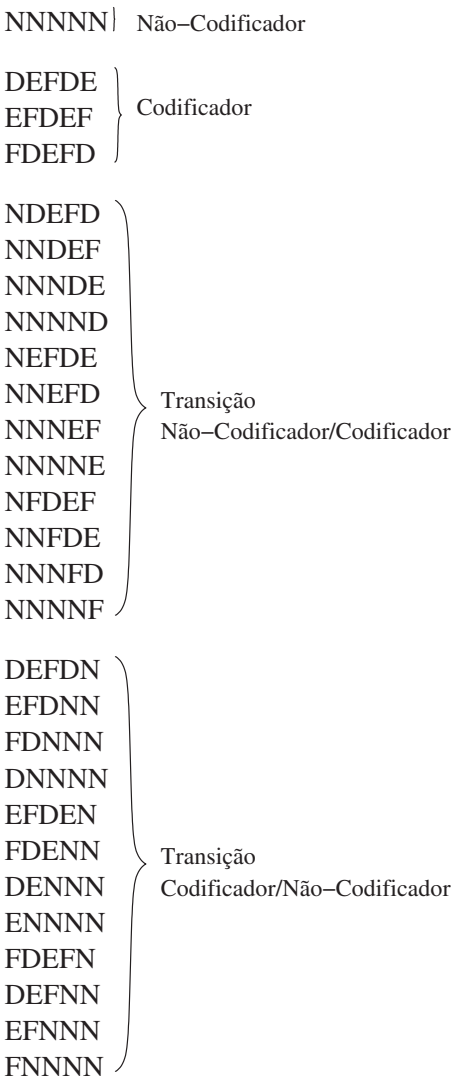

Figura 5.2: 5-eventos considerados para uma 5-tupla.

De maneira análoga, a hipótese da independência de bases nas regiões não-codificadoras permite decompor $\Pr(\text{ATCG}|\text{NNNN})$ em $\Pr(\text{A}|\text{N}) \times \Pr(\text{T}|\text{N}) \times \Pr(\text{C}|\text{N}) \times \Pr(\text{G}|\text{N})$. Finalmente, supomos que as fronteiras entre regiões codificadoras (C) e não-codificadoras (N) são aleatórias e independentes dos conteúdos, o que nos permite quebrar eventos nas fronteiras C-N e N-C, portanto $\Pr(\text{TTACG}|\text{NNDEF})$ pode ser decomposto em $\Pr(\text{T}|\text{N}) \times \Pr(\text{T}|\text{N}) \times \Pr(\text{ACG}|\text{DEF})$.

Em suma, podemos decompor $\Pr(\alpha|\beta)$, num certo número v de fatores:

$$\Pr(\alpha|\beta) = \Pr(\alpha_1|\beta_1) \times \Pr(\alpha_2|\beta_2) \times \cdots \times \Pr(\alpha_v|\beta_v) \tag{5.3}$$

onde, $\alpha = \alpha_1, \alpha_2, \ldots, \alpha_v$, $\beta = \beta_1, \beta_2, \ldots, \beta_v$ e $|\alpha_i| = |\beta_i| \leq 3$ para todo $i = 1, \ldots, v$. Desta forma, cada termo $\Pr(\alpha_i|\beta_i)$ é obtido por tabelas de tamanho máximo igual a 3, conforme a seção 5.3.

5.2.5.2 Cálculo de $\Pr(\beta)$

Para o cálculo da probabilidade *a priori* $\Pr(\beta)$, supomos que os comprimentos dos trechos C (codificador - D,E,F) e N (não-codificador) têm alguma distribuição aleatória com mínimo t_{min}, e médias t_c e t_n respectivamente. Portanto, a probabilidade de uma base arbitrária ter rótulo C ou N é dada da seguinte forma: considerando uma longa seqüência de DNA onde temos todas as k-tuplas e supondo que o tamanho mínimo das regiões C ou N é $k-1$, uma vez que nenhuma tupla contém mais de uma transição C-N ou N-C. Como o tamanho médio das regiões C e N é t_c e t_n haverá uma transição C-N e uma transição N-C a cada $(t_c + t_n)$ bases. Então, a fração de k-tuplas de cada tipo é:

$$\begin{aligned}
\text{C-N Misturados} &: \frac{(k-1)}{(t_c+t_n)} \\
\text{N-C Misturados} &: \frac{(k-1)}{t_c+t_n} \\
\text{N puro} &: \frac{(t_n-k+1)}{(t_c+t_n)} \\
\text{C puro} &: \frac{(t_c-k+1)}{(t_c+t_n)}
\end{aligned}$$

Para cada possível padrão “C puro” há três eventos, correspondendo às três possíveis fases. Desta forma a probabilidade de cada evento é igual à fração de “C puro” dividida por 3. O mesmo vale para cada padrão misturado C-N ou N-C.

5.3 Treinamento

Antes de aplicar a classificação, é necessário determinar para cada k-tupla $\alpha \in \Gamma^k$ e cada evento $\beta \in \Upsilon^k$, a probabilidade $\Pr(\alpha|\beta)$ de k nucleotídeos consecutivos de DNA terem bases $b_1, \ldots, b_k$ e rótulos $e_1 \ldots e_k$, onde k é um parâmetro dado. Isso é geralmente feito pela análise de amostras de regiões codificadoras e não-codificadoras conhecidas.

Por simplificação, vamos supor que os dados de amostras consistem de uma única seqüência $s = s_1, \ldots, s_n$; e que cada base s_i tem uma classificação conhecida $e_i \in \Upsilon$.

Primeiro, calculamos o número total $\#(\alpha \wedge \beta)$ de ocorrências da tupla α na seqüência s com rótulos do k-evento β, para cada $\alpha \in \Gamma^k$ e cada $\beta \in \Upsilon^k$ — produzindo deste modo, uma tabela com $4^k \times 4^k$ entradas.

Intuitivamente, poderíamos estimar a probabilidade $\Pr(\alpha|\beta)$ pela fórmula:

$$\Pr(\alpha|\beta) = \frac{\#(\alpha \wedge \beta)}{\#(\beta)} \qquad \text{onde} \qquad \#(\beta) = \sum_{\alpha \in \Gamma^k} \#(\alpha \wedge \beta) \tag{5.4}$$

Contudo, se a amostra tem tamanho relativamente pequeno, certos pares (α, β) podem não aparecer em s por mero azar. Nesse caso, a fórmula (5.4) atribuiria $\Pr(\alpha|\beta) = 0$ apesar de que, na realidade, sua probabilidade é maior que 0. Ou seja, $\#(\alpha \wedge \beta) = 0$ deve ser entendido como "$(\alpha \wedge \beta)$ é raro demais para ocorrer na amostra" e não "$(\alpha \wedge \beta)$ é impossível". Em vista destas considerações usamos a estimativa:

$$\Pr(\alpha|\beta) = \frac{\#(\alpha \wedge \beta) + 1}{\#(\beta) + 4^k} \tag{5.5}$$

Vale notar que tanto a equação (5.4) quanto (5.5) tendem para o mesmos limites quando o número de bases da amostra tende para infinito.

Capítulo 6

Testes e resultados

Neste capítulo apresentamos os resultados da implementação do método proposto aplicados a um conjunto de seqüências de teste. Na seção 6.1, fornecemos uma breve descrição sobre os programas implementados, na seção 6.3, detalhamos as seqüências utilizadas no conjunto de testes e na seção 6.4, exibimos os resultados experimentais obtidos.

6.1 O programa de classificação

O método de classificação proposto é implementado por um programa chamado `Pre`, que faz uma classificação probabilística de cada base de uma seqüência de DNA fornecida nos eventos D, E, F e N. Opcionalmente, o programa compara esse resultado com uma rotulação "oficial" da seqüência e imprime uma avaliação da classificação realizada.

Parâmetros de entrada Devem ser fornecidos como entrada do programa `Pre` o valor k (tamanho da tupla), o tamanho médio de uma região codificadora (t_c), o tamanho médio de uma região não-codificadora (t_n) e três arquivos com as tabelas de probabilidades $Pr(\alpha|\beta)$ para $|\beta| = \{1, 2, 3\}$. Além destes, o `Pre` necessita de um arquivo com todos os k-eventos β a considerar, e suas fatorações em 1, 2 e 3 eventos. Finalmente, devem ser informados os nomes dos arquivos contendo a seqüência de nucleotídeos a ser rotulada (seqüência alvo), e sua rotulação "oficial", se disponível.

Saída O programa `Pre` produz como saída um arquivo com as probabilidades $Pr^*[i, \epsilon]$ para $\epsilon =$ D,E,F,N de cada base b_i da seqüência alvo como descrita na seção 5.2. Um exemplo pode ser visto na figura 6.1.

position	b	Pr(D)	Pr(E)	Pr(F)	Pr(N)	labs	Pr(C)	Pr(N)	codn	Pr(D)	Pr(E)	Pr(F)	phas
0	G	0.000000	0.000000	0.000000	0.000000	?:N?	0.000000	0.000000	?:N?	0.000000	0.000000	0.000000	?:??
1	T	0.045031	0.022666	0.016353	+0.915949	N:N·	0.084051	+0.915949	N:N·	+0.535767	0.269672	0.194561	D:??
2	T	0.022285	0.031593	0.016261	+0.929861	N:N·	0.070139	+0.929861	N:N·	0.317731	+0.450430	0.231840	E:??
3	T	0.019775	0.028741	0.040284	+0.911199	N:N·	0.088801	+0.911199	N:N·	0.222693	0.323657	+0.453650	F:??
4	G	0.060902	0.004757	0.076245	+0.858097	N:N·	0.141903	+0.858097	N:N·	0.429178	0.033521	+0.537301	F:??
5	A	0.052449	0.095129	0.027956	+0.824466	N:N·	0.175534	+0.824466	N:N·	0.298794	+0.541941	0.159265	E:??
6	C	0.038460	0.031354	0.039316	+0.890869	N:N·	0.109131	+0.890869	N:N·	0.352422	0.287310	+0.360269	F:??
7	T	0.038558	0.142978	0.057687	+0.760778	N:N·	0.239222	+0.760778	N:N·	0.161179	+0.597678	0.241144	E:??
8	G	0.031344	0.023159	0.049741	+0.895755	N:N·	0.104245	+0.895755	N:N·	0.300680	0.222162	+0.477159	F:??
9	T	0.028442	0.016047	0.026539	+0.928973	N:N·	0.071027	+0.928973	N:N·	+0.400437	0.225923	0.373640	D:??
10	A	0.028566	0.025257	0.007451	+0.938727	N:N·	0.061273	+0.938727	N:N·	+0.466203	0.412195	0.121601	D:??
11	T	0.016182	0.061568	0.045484	+0.876767	N:N·	0.123233	+0.876767	N:N·	0.131311	+0.499603	0.369086	E:??
12	G	0.045012	0.046599	0.092465	+0.815923	N:N·	0.184077	+0.815923	N:N·	0.244530	0.253153	+0.502318	F:??
13	C	0.071759	0.043493	0.068200	+0.816549	N:N·	0.183451	+0.816549	N:N·	+0.391161	0.237080	0.371759	D:??
14	A	0.089312	0.030447	0.032907	+0.847333	N:N·	0.152667	+0.847333	N:N·	+0.585012	0.199437	0.215551	D:??
15	A	0.033901	0.048369	0.034534	+0.883197	N:N·	0.116803	+0.883197	N:N·	0.290239	+0.414102	0.295658	E:??

Figura 6.1: Trecho de um arquivo de saída do `Pre` com a classificação para uma seqüência de bases.

Neste arquivo, cada linha refere-se a uma base da seqüência de DNA alvo e possui o formato:

```
''{i} {bi} {prDEFN} {emDEFN}  {prCN} {emCN}  {prDEF} {emDEF}''
```

onde o campo $\{i\}$ indica a posição da base $\{bi\}$ na seqüência, e o campo {prDEFN} consiste das quatro probabilidades computadas $\Pr(e_i = \epsilon)$, onde e_i é o verdadeiro rótulo (desconhecido) de b_i e $\epsilon \in \Upsilon$. O campo {emDEFN} resume algumas informações de {prDEFN} e possui o formato X:YZ, onde X é o rótulo em {prDEFN} que possui maior probabilidade (e que também é marcado com um "+" na coluna {prDEFN}), Y é o rótulo oficial (gabarito) de $\{bi\}$ e Z recebe uma marca — "." se X e Y forem iguais, e "*" se forem diferentes.

O campo {prCN} é similar ao {prDEFN}, mas desconsiderando a fase de leitura dos códons. Assim, as probabilidades dos rótulos D, E e F são condensadas na probabilidade do rótulo C. O campo {emCN}, do mesmo modo que {emDEFN}, resume as informações de {prCN}.

O campo {prDEF}, ao contrário, considera apenas a fase de leitura, supondo que a base b_i é codificadora. Este é obtido a partir das probabilidades do campo {prDEFN}, descartando-se a probabilidade Pr(N) e normalizando-se as probabilidades restantes para somar um. Já o conteúdo de Z no campo {emDEF} recebe "." se X = Y, "+" se X situa-se uma posição a mais no códon que Y (E:D, F:E ou D:F) e "-" se X encontra-se uma posição a menos (F:D, D:E ou E:F).

Campos com informação não disponível podem ser marcados com "?". Isso acontece, por exemplo, nas primeiras $(t-1)$ e nas últimas $(t-1)$ bases da seqüência, onde as probabilidades $\Pr^*$ são todas nulas, significando que não foram computadas. A marca "?" é também usada quando o rótulo é desconhecido ou nulo, por exemplo no campo {emDEF} quando o rótulo oficial de $\{bi\}$ é "N".

Quando a rotulação oficial é fornecida, o programa `Pre` também gera um outro arquivo, chamado de *escore*, com a avaliação da rotulação. Um exemplo desse arquivo pode ser visto na figura 6.2.

```
          Est(D)    Est(E)    Est(F)    Est(N)  Accu.
       --------- --------- --------- ---------  -----
Pr(D)   1037.80     29.16     23.68    586.36  0.618
Pr(E)     23.30   1034.34     29.62    589.74  0.616
Pr(F)     28.75     23.66   1032.05    592.54  0.615
Pr(N)    536.50    539.09    542.67  23286.74  0.935

[DEFN]       30000 bases       29936 considered     26390.93 matched  accuracy = 0.881

          Est(C)    Est(N)  Accu.
       --------- ---------  -----
Pr(C)   3262.37   1768.63  0.648
Pr(N)   1618.26  23286.74  0.935

[CN]         30000 bases       29936 considered     26549.10 matched  accuracy = 0.887

          Est(D)    Est(E)    Est(F)  Accu.
       --------- --------- ---------  -----
Pr(D)   1488.58     92.55     95.86  0.887
Pr(E)     96.04   1486.24     94.72  0.885
Pr(F)     94.40     97.43   1485.17  0.885

[DEF]        30000 bases        5031 considered      4460.00 matched  accuracy = 0.886
```

Figura 6.2: Trecho de um arquivo de escore, saída do `Pre`.

O arquivo de escore é constituído por 3 tabelas. A primeira tabela, que chamaremos de $\{$scrDEFN$\}$ ou simplesmente S descreve a qualidade da previsão $\Pr^*[i, \epsilon]$, isto é, do campo $\{$prDEFN$\}$. Para cada $\epsilon, \epsilon' \in \Upsilon$, a célula $S[\epsilon, \epsilon']$ é a soma de $\Pr^*[i, \epsilon]$ sobre todas as bases b_i cujo rótulo oficial é ϵ'.

Uma boa classificação realizada pelo preditor gera uma matriz onde os maiores valores das probabilidades de cada linha pertencem à diagonal principal.

A coluna "Accu." da tabela informa o escore do programa separado para cada rótulo. Na linha ϵ, esse escore é o valor de $S[k, k]$ dividido pela soma dos elementos da linha (que é o número de ocorrências do rótulo ϵ_k na rotulação oficial). Quanto maior o seu valor, melhor é a precisão do programa para bases com esse rótulo.

A linha "[DEFN]" logo abaixo da tabela resume algumas informações de $\{$scrDEFN$\}$: a quantidade de bases existentes na seqüência (*bases*), o número de bases efetivamente classificadas pelo preditor (*considered*), a soma da diagonal principal (*matched*) e o escore, que é o quociente destas duas últimas (*accuracy*).

Esta tabela é seguida pelas tabelas $\{$scrCN$\}$ (avaliação de $\{$prCN$\}$) e $\{$scrDEF$\}$ (avaliação de $\{$prDEF$\}$), calculadas do mesmo modo que $\{$scrDEFN$\}$.

6.2 Programas auxiliares

Existem ainda outros programas que foram implementados, explicados a seguir.

1. `Tab`: O programa `Tab` extrai dos dados de treinamento todos os pares (α, β), onde α é uma k-tupla (conjunto de k bases consecutivas), e β o seu correspondente k-evento (conjunto de k rótulos consecutivos). Para cada par (α, β) diferente, o programa calcula sua freqüência de ocorrência e a probabilidade condicional $\Pr(\alpha|\beta)$, conforme descrita em 5.3. O valor de k nas tuplas calculadas por `Tab` são 1, 2 e 3 respectivamente, pois consideramos que todas as outras probabilidades podem ser obtidas pela fatoração do k-evento em eventos de tamanho 1, 2 e 3.

   Este programa tem dois parâmetros de entrada:

   (a) um inteiro k, que é o tamanho da janela (assumido como 1, 2 e 3);

   (b) um arquivo onde cada linha contém os nomes de pares de arquivos. Em cada par, o primeiro arquivo contém as bases de uma seqüência de DNA (letras A, T , C e G) e o segundo, seus respectivos rótulos.

   A saída do `Tab` é um arquivo com todos os pares (α, β) e as probabilidades $\Pr(\alpha|\beta)$.

2. `Extract-bas-lab-trn`: programa que extrai três arquivos da base de dados genômica, um contendo as seqüências de bases (possui o mesmo nome da seqüência do banco de dados porém com a extensão .bas), outro com os rótulos correspondentes a cada uma destas bases (com a extensão .lab) e o último com as regiões codificadoras divididas em códons pareados com o respectivo aminoácido (com a extensão .trn). Este último arquivo é utilizado para checar se a seqüência codificadora pertencente à base de dados é equivalente ao seu conjunto de aminoácidos informados.

3. `Enum-valid-events`: programa que gera todos os eventos válidos com as fatorações a serem usadas no cálculo das probabilidades Pr(k-tupla|k-evento) e as probabilidades *a priori* (Pr(k-evento)), segundo descrito na seção 5.2.5.

Os programas `Pre` e `Tab` foram implementados usando a linguagem C, seguindo o padrão C ANSI, com o compilador gcc versão 4.0.0 em ambiente Linux (distribuição Fedora Core 4). Os programas `Extract-bas-lab-trn` e `Enum-valid-events` foram implementados usando a linguagem Awk GNU versão 3.1.4.

6.3 Conjunto de seqüências de teste

As seqüências que usamos para testar o programa são uma coleção de trechos do genoma humano coletadas por Martin Reese [1]. Esta coleção consiste de 7 conjuntos com um total de 462 arquivos, cada um deles contendo a seqüência de bases de um gene (ou trecho de gene). Em cada arquivo há um cabeçalho com informações que permitem identificar as regiões codificadoras ativas (éxons) da seqüência.

Por segurança, eliminamos dessa coleção todos os arquivos onde os éxons foram identificados de forma não experimental — de modo geral, por outros programas classificadores. Além disso, para simplificar o tratamento dos dados, eliminamos também seqüências com éxons na fita complementar.

Foram eliminadas ainda seqüências com nucleotídeos não-identificados ou parcialmente identificados. Estes nucleotídeos são representados nos arquivos originais por outras letras além de A, C, T e G. Finalmente, por segurança desconsideramos arquivos com múltiplos campos "sources" (campos que indicam em que posição da fita se situa o trecho estudado).

Depois desta filtragem, restaram 448 arquivos, 64 em cada conjunto numerado de 0 a 6, totalizando 5.258.181 bases; sendo que 439.761 destas bases estão situadas em regiões codificadoras ativas, totalizando 146.587 códons.

6.4 Experimentos realizados

Realizamos alguns experimentos computacionais para avaliação de nosso método implementado no reconhecimento das regiões codificadoras e da correta fase de leitura. Os testes foram realizados em uma máquina com processador Pentium IV 3.00GHz e 1Gb de memória RAM.

Nos testes realizados, os tamanhos das janelas de leitura considerados são: 1 e $2^r + 1$ para r variando de 0 a 6. Não testamos janelas com $k = 257$ e maiores porque excederia o tamanho médio dos éxons (137 pares de bases [24]).

6.4.1 Seqüência sintética Y

Para verificar a habilidade do `Pre` em reconhecer as características estatísticas consideradas, geramos uma seqüência aleatória de DNA de acordo com o nosso modelo estocástico. Essa seqüência, chamada de Y, foi utilizada para testar o programa. As bases das regiões não-codificadoras de Y foram escolhidas independentemente de acordo com as probabilidades da

tabela 5.3; e os códons das regiões codificadoras foram escolhidas independentemente de acordo com as probabilidades apresentadas na tabela 5.2. A seqüência Y gerada tinha 45.000 bases, com regiões codificadoras e não-codificadoras com distribuição exponencial de tamanhos, definidos com tamanho mínimo $t_{min} = 100$, e tamanhos médios $t_c = 150$ e $t_n = 900$. Todas as bases das regiões não-codificadoras receberam rótulos "N", e as bases das regiões codificadoras receberam os rótulos "D", "E" e "F" de acordo com sua posição dentro do códon. Para o treinamento usamos todos os conjuntos de seqüências de regiões codificadoras e não-codificadoras do genoma humano, descritas na seção anterior.

As tabelas 6.1, 6.2 e 6.3 resumem as probabilidades encontradas na execução do `Pre` (arquivo de escore) para a seqüência Y, considerando uma janela de tamanho igual a 65.

	Est(D)	**Est(E)**	**Est(F)**	**Est(N)**	**Accu.**
Pr(D)	**1543,76**	26,05	49,68	600,51	0,695
Pr(E)	51,16	**1539,69**	26,75	602,40	0,693
Pr(F)	26,04	51,66	**1533,30**	609,00	0,690
Pr(N)	733,89	736,66	741,33	**36064,12**	0,942

Tabela 6.1: Resumo das probabilidades encontradas para as bases da seqüência Y com uma janela de tamanho 65.

	Est(C)	**Est(N)**	**Accu.**
Pr(C)	**4848,09**	1811,91	0,728
Pr(N)	2211,88	**36064,12**	0,942

Tabela 6.2: Resumo das probabilidades encontradas para a seqüência Y com uma janela de tamanho 65 sendo Est(C) = Est(D) + Est(E) + Est(F) e Pr(C) = Pr(D) + Pr(E) + Pr(F).

	Est(D)	**Est(E)**	**Est(F)**	**Accu.**
Pr(D)	**2032,58**	78,03	109,40	0,915
Pr(E)	112,60	**2028,51**	78,89	0,913
Pr(F)	76,33	114,73	**2028,94**	0,913

Tabela 6.3: Resumo das probabilidades encontradas para a predição da fase de leitura nas regiões codificadoras da seqüência Y com uma janela de tamanho 65.

De modo geral, como existem mais regiões não-codificadoras do que regiões codificadoras no genoma humano, a probabilidade de uma base estar localizada em uma região codificadora obtida no treinamento é pequena. Essa diferença ainda não é superada por uma janela de tamanho 65, como exibida na tabela 6.1.

Apresentamos na figura 6.3 alguns gráficos para ilustrar o comportamento do programa com diversos tamanhos de janela. Para isto, tomamos o trecho da seqüência Y consistindo das bases 0 a 499. Esse trecho possui duas regiões não-codificadoras, separadas por uma região

codificadora. A classificação oficial de cada base é representada pela linha pontilhada (O), onde o valor 0 é atribuído às bases situadas nas regiões não-codificadoras e 1 às codificadoras. A probabilidade estimada Pr(C) do arquivo de classificação encontrada pelo preditor para cada base desse trecho de Y é representada pela linha cheia (C).

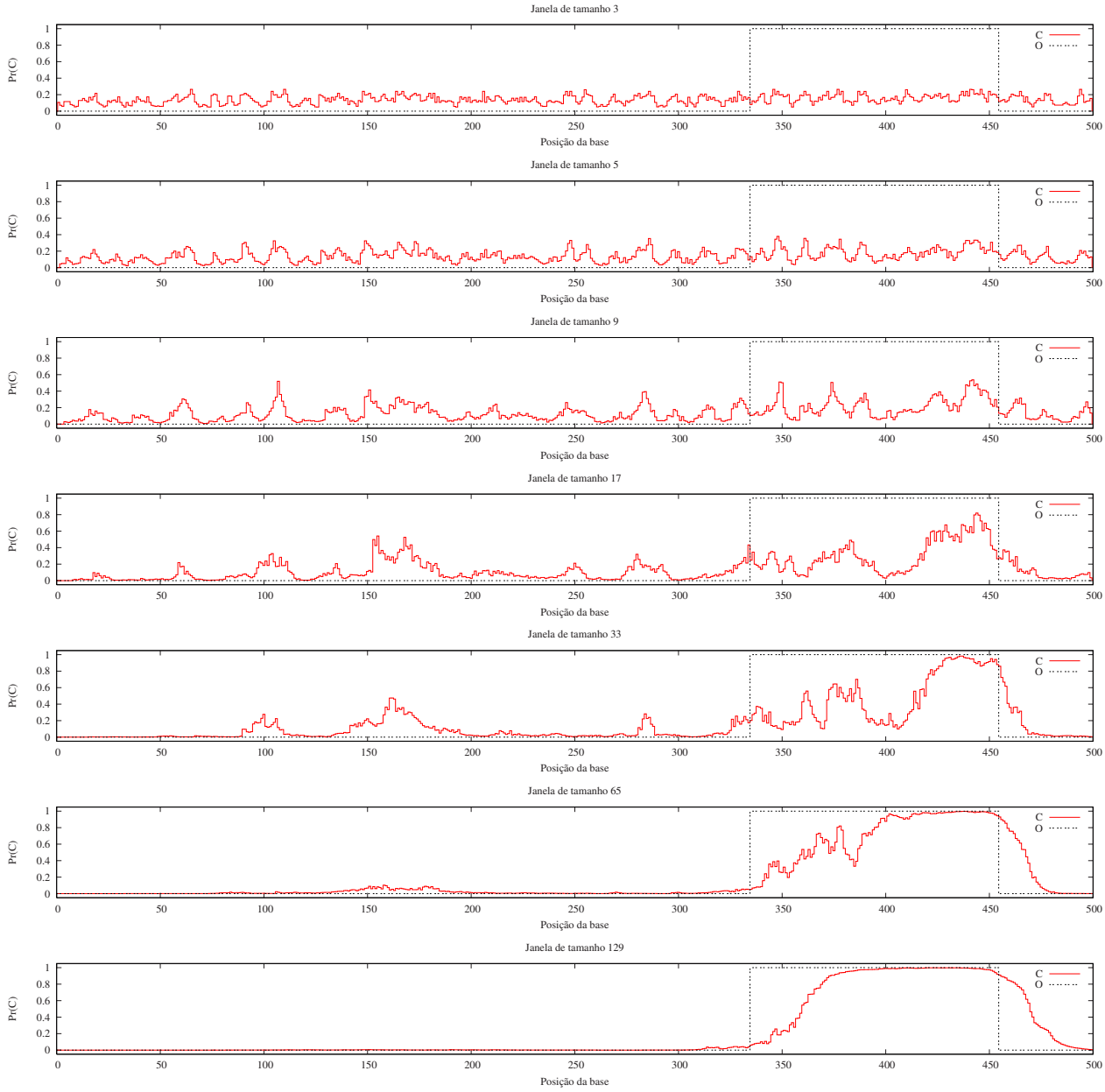

Figura 6.3: Gráficos da execução do `Pre` em vários tamanhos de janela considerados para um trecho da seqüência Y.

Por estes gráficos, observamos que as probabilidades Pr(C) em janelas pequenas não fornecem muita informação para o reconhecimento das regiões codificadoras. Porém, esta informação aumenta a medida que o tamanho da janela cresce. Isso pode ser melhor visto na figura 6.4, que ilustra os valores da média de Pr(C) (campo “Accu.” da linha Pr(C) nos arquivos de escore) nas

execuções do programa para todas as janelas consideradas. Note que, como previsto na seção 4.4.2, a medida que aumenta o tamanho da janela de leitura, aumenta também a probabilidade de uma base ser corretamente classificada como codificadora. Essa precisão chega a 0,839 para uma janela de tamanho igual a 129.

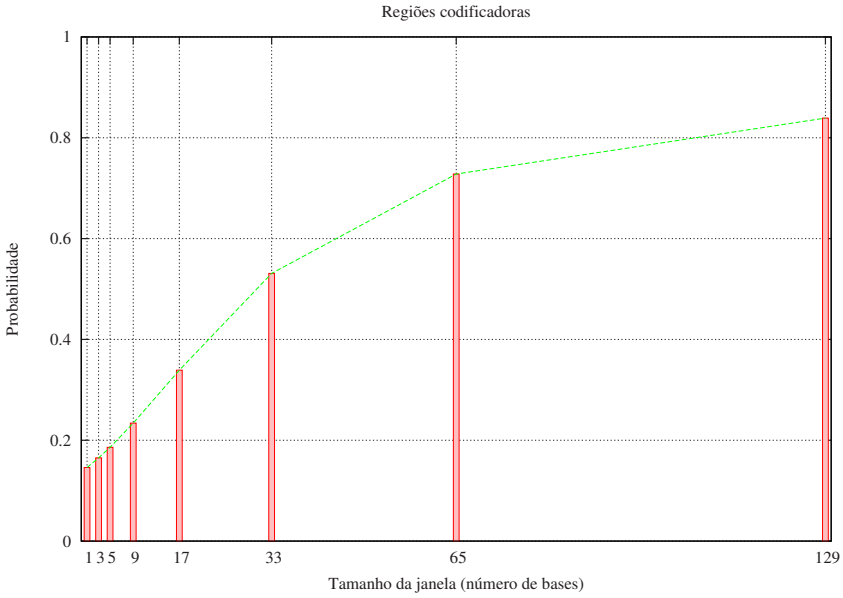

Figura 6.4: Identificação das regiões codificadoras na seqüência Y.

Na figura 6.5 são exibidas as médias de todos os elementos exibidos nos arquivos de escore no campo "[DEF]". Cada campo deste, chamado de "*accuracy*", exibe a soma dos elementos da diagonal principal dividida pelo total de bases consideradas no arquivo (excluindo as $t-1$ bases no começo e as $t-1$ bases no final da seqüência). Estes valores de probabilidade encontrados representam a precisão do programa para a identificação da correta fase de leitura nas regiões codificadoras da seqüência Y. Novamente neste gráfico, observamos um aumento das médias de "*accuracy*", chegando a 0.982, quando o tamanho da janela é de 129.

Temos ainda que o tempo de CPU gasto em segundos para classificar a seqüência Y em cada uma das janelas é visto na tabela 6.4. Observamos um grande acrescimo no tempo a medida que o tamanho de k aumenta.

6.4.2 Conjunto de seqüências do genoma humano

Para observar o comportamento do programa e avaliar as probabilidades fornecidas para um conjunto de dados reais, treinamos e classificamos os nucleotídeos do genoma humano, descritos anteriormente (seção 6.3). Separamos os 7 conjuntos da base de dados em dois grupos e treinamos com o primeiro grupo, composto pelos conjuntos 0, 1, 2 e 3, com 3.208.731 bases.

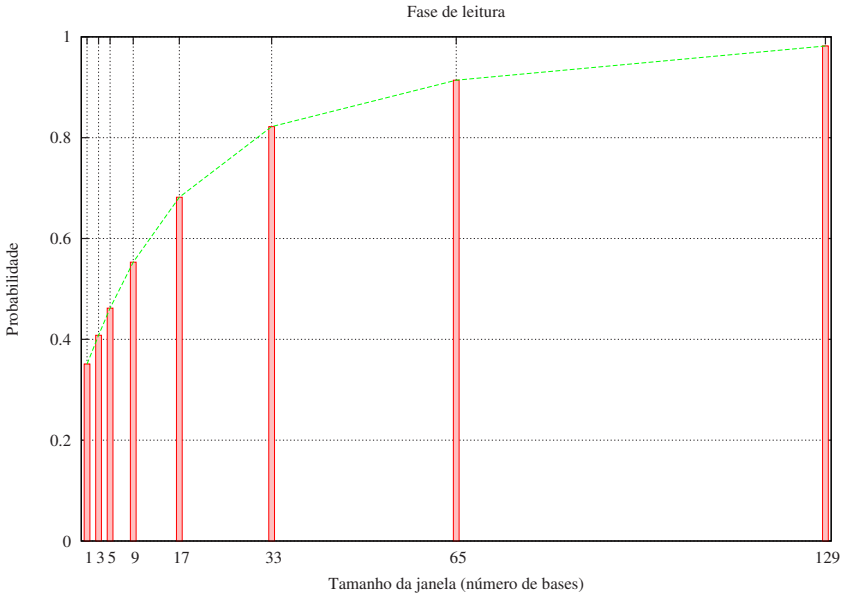

Figura 6.5: Identificação da correta fase de leitura para as regiões codificadoras da seqüência Y.

k	tempo total(s)
1	0,544
3	0,688
5	0,915
9	1,653
17	4,022
33	12,663
65	46,553
129	179,182

Tabela 6.4: Tempo de execução do programa `Pre` para classificar a seqüência Y.

O procedimento de classificação foi realizado com o segundo grupo, composto pelos conjuntos 4, 5 e 6, com 2.049.450 bases.

Para os testes, adotamos $t_c = 150$ e $t_n = 900$. O valor correto desses parâmetros deve ser o tamanho médio de cada tipo de região que se espera encontrar na seqüência a ser analisada. Se o DNA analisado é de uma bactéria, devemos usar $t_c > t_n$; se é de um humano, por exemplo, devemos usar $t_c < t_n$, já que sabemos que no genoma humano as regiões não-codificadoras tem tamanho médio muito maior do que os das regiões codificadoras. Esses parâmetros influenciam o resultado através do $\Pr(\beta)$, especialmente na razão $t_c/(t_c + t_n)$ (porcentagem esperada de bases codificadoras na seqüência analisada) e no valor absoluto de $(t_c + t_n)$ (que determina $\Pr(\beta)$ para eventos mistos).

6.4.2.1 Uma seqüência de estudo

Uma seqüência do segundo grupo de arquivos (grupo de classificação) será analisada nesta subseção para ilustrar o desempenho do `Pre` para um arquivo de seqüências de bases real. A seqüência, que foi escolhida aleatoriamente, está no arquivo chamado de D49493 e possui um total de 17.286 bases. As tabelas do arquivo de escore, resultado da execução de `Pre` para a seqüência do arquivo D49493, com uma janela de tamanho 65, são exibidas nas tabelas 6.5, 6.6 e 6.7.

	Est(D)	**Est(E)**	**Est(F)**	**Est(N)**	**Accu.**
Pr(D)	**392,02**	14,23	14,45	58,30	0,814
Pr(E)	15,31	**392,41**	13,52	57,76	0,815
Pr(F)	13,56	16,99	**387,90**	60,55	0,805
Pr(N)	1966,01	1945,88	1939,99	**9933,13**	0,629

Tabela 6.5: Resumo das probabilidades encontradas para a seqüência D49493 com uma janela de tamanho 65.

	Est(C)	**Est(N)**	**Accu.**
Pr(C)	**1260,40**	176,60	0,877
Pr(N)	5851,87	**9933,13**	0,629

Tabela 6.6: Resumo das probabilidades encontradas para a seqüência D49493 com uma janela de tamanho 65 sendo Est(C) = Est(D) + Est(E) + Est(F) e Pr(C) = Pr(D) + Pr(E) + Pr(F).

	Est(D)	**Est(E)**	**Est(F)**	**Accu.**
Pr(D)	**440,36**	19,95	18,69	0,916
Pr(E)	19,61	**440,34**	19,05	0,916
Pr(F)	19,20	21,84	**437,96**	0,911

Tabela 6.7: Resumo das probabilidades encontradas para a predição da fase de leitura nas regiões codificadoras da seqüência D49493 com uma janela de tamanho 65.

Como podemos observar, os valores do campo “Accu.” (tabelas 6.5 e 6.6) para os rótulos das regiões codificadoras na seqüência testada são relativamente altos se comparados com os obtidos pela seqüência Y, isso ocorre possivelmente por vários motivos. Um deles é devido a presença de códons com altas taxas de preferência nas regiões codificadoras (códon que aparecem muito mais nas regiões codificadoras — elevada freqüência de ocorrência), exibindo bem as propriedades estatísticas capturadas dos dados de treinamento. Outro motivo é a ausência em D49493 de regiões codificadoras de tamanho pequeno, respeitando a restrição de somente uma quebra N-C ou C-N. Já os valores de “Accu.” para Pr(N) diminuiram drasticamente. Isso ocorre pois como apresentado na seção 5.2.3.2, os nucleotídeos pertencentes às regiões não-codificadoras apresentam relações de dependências, provavelmente devido à presença de

regiões codificadoras fósseis em trechos de suas seqüências. Tal problema se resolveria apenas com a adoção de um modelo mais realista para essas regiões.

As probabilidades encontradas na tabela 6.7 do mesmo modo que as apresentadas para a seqüência Y, indicam uma boa taxa de reconhecimento da fase de leitura.

A seqüência apresenta outra propriedade, similar a Y, de modo que à medida que a janela de leitura aumenta, o valor de Pr(D) aumenta também. Podemos acompanhar tal fato no gráfico visto na figura 6.6 para um trecho da sequência D49493 de 500 bases.

6.4.2.2 Todas as seqüências do grupo de classificação

Analisaremos agora a execução do `Pre` para todas as seqüências do grupo de classificação. Os valores das probabilidades do campo "Accu." de Pr(C) nas tabelas de escore para todos os arquivos do conjunto de testes podem ser vistos nos gráficos da figura 6.7. Os valores das probabilidades encontrados, representadas por pontos, são mostrados em cada tamanho de janela, cada ponto representa o valor "Accu." de Pr(C) de um arquivo (ou seqüência) e cada reta representa o valor médio de todas as probabilidades exibidas em cada gráfico.

Como podemos observar, uma grande parcela das probabilidades que em janelas menores localizavam-se na parte inferior dos gráficos, passam a se concentrar na parte superior quando o tamanho da janela cresce. Quase todos os valores de probabilidade sofrem algum aumento, exceto os que pertencem às seqüências com regiões codificadoras de tamanho pequeno. Como em nosso modelo de DNA não existe mais de uma quebra N-C ou C-N, o tamanho mínimo das regiões codificadoras aqui identificadas é dado pelo tamanho da janela analisada menos 1. Em janelas de tamanho 129 por exemplo, regiões codificadoras pequenas (com menos de 128 bases) provavelmente não serão corretamente reconhecidas.

O aumento das probabilidades acompanhando o aumento no tamanho da janela podem ser melhor vistos pelo gráfico da figura 6.8, onde apresentamos o valor médio de todas as probabilidades exibidas nos gráficos da figura 6.6 e representadas por uma linha contínua (média). Estes valores, fornecem uma média das probabilidades que uma base com rótulo "C" receberia ao ser classificada como "C". Este valor chegou a 0,759 em janelas de 129.

Finalmente, na figura 6.9 analisamos o comportamento do programa no reconhecimento da correta fase de leitura para as bases do genoma humano. Cada ponto representa a média de todos os campos da tabela {scrDEF} no arquivo de escore em cada seqüência utilizada na classificação. Cada reta representa a média do conjunto de pontos para cada arquivo.

Observamos nas probabilidades do genoma humano o mesmo comportamento das proba-

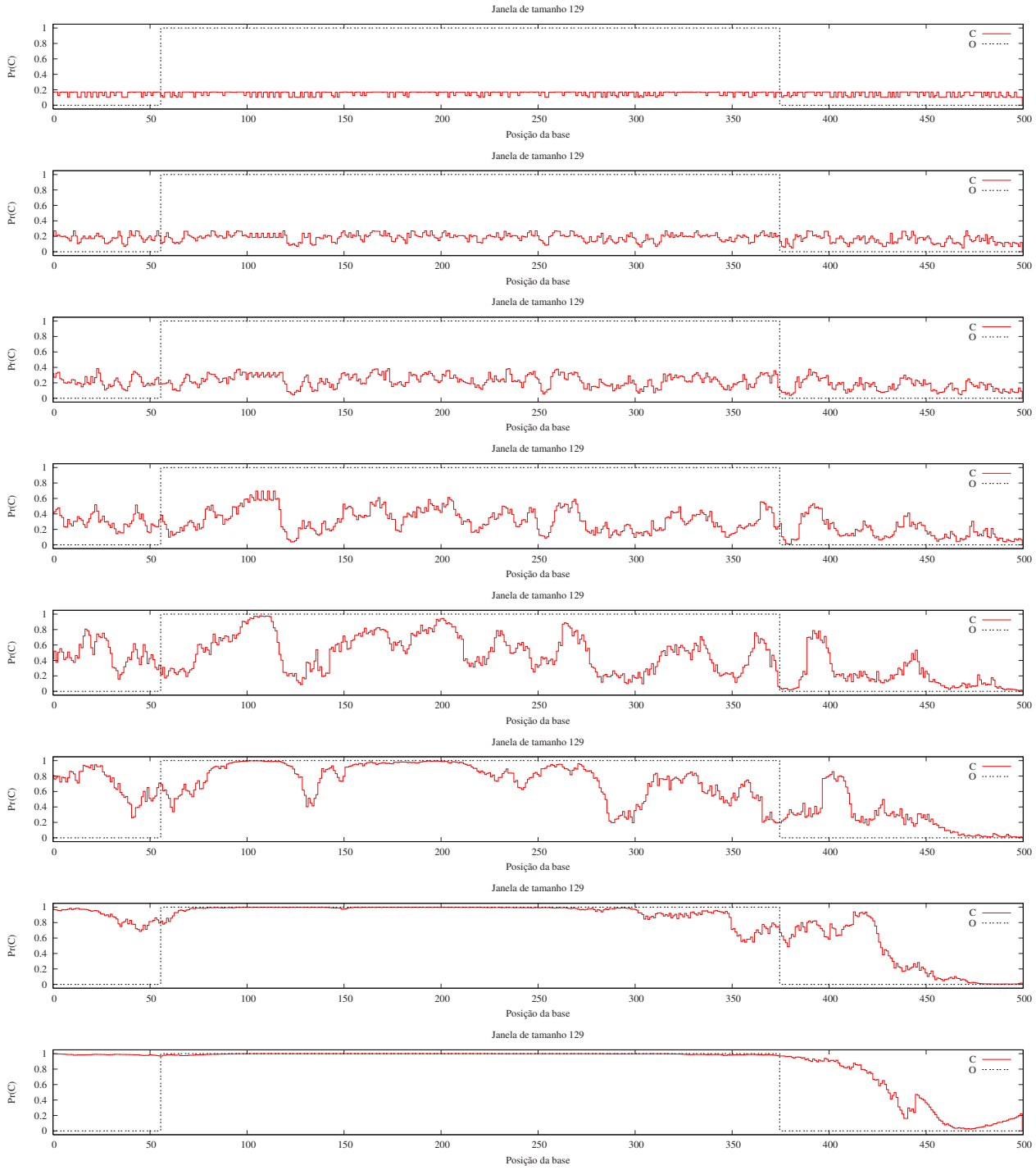

Figura 6.6: Gráficos da execução do `Pre` em vários tamanhos de janela considerados para um trecho da seqüência D49493.

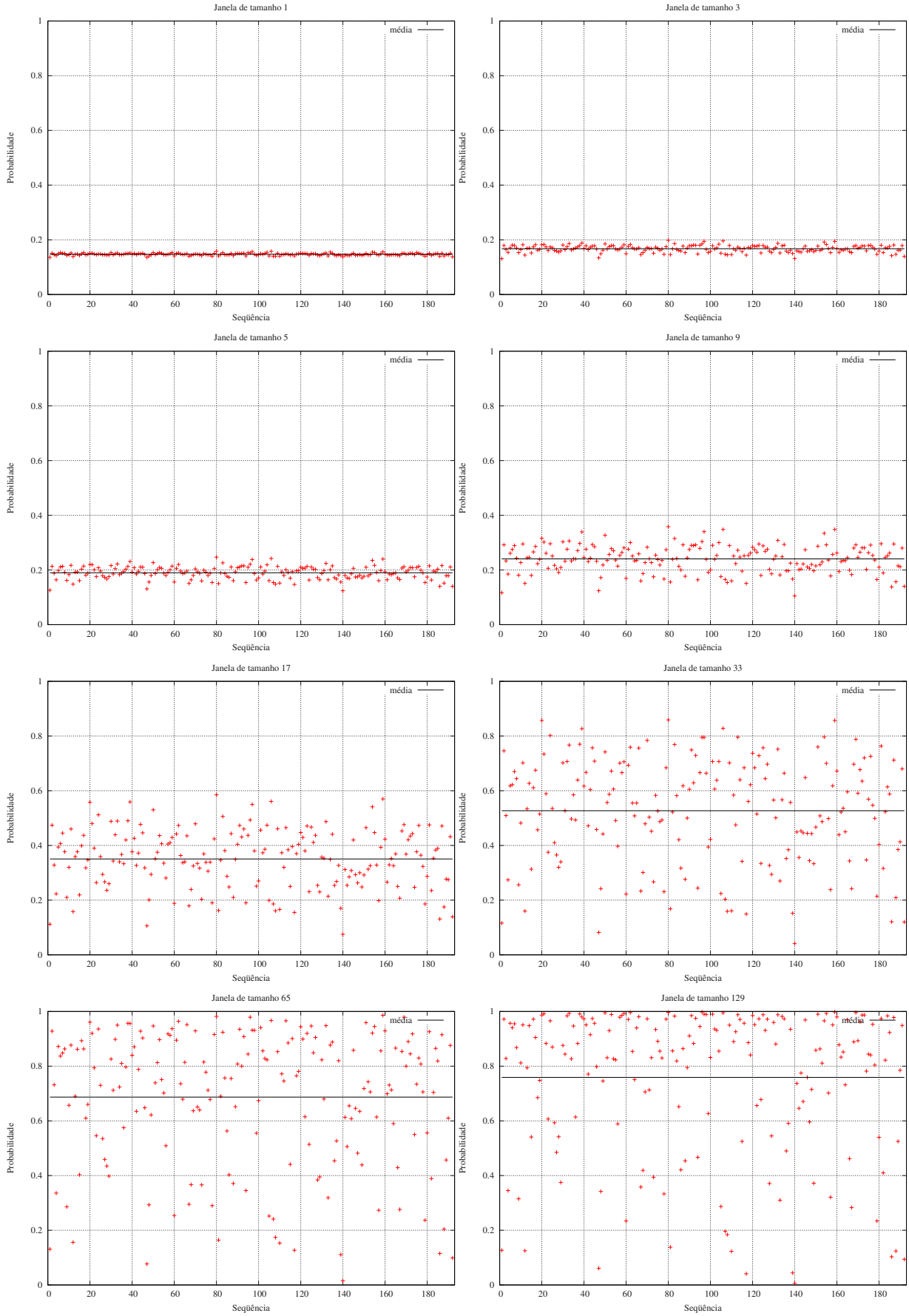

Figura 6.7: Probabilidades do campo Accu. de Pr(C) na tabela {scrCN} para as diversas janelas em todos os arquivos de escore.

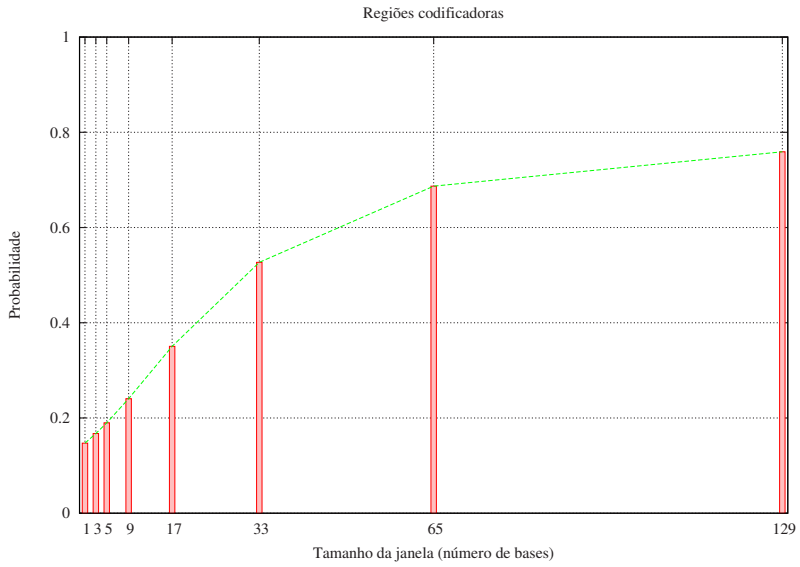

Figura 6.8: Identificação das regiões codificadoras no genoma humano.

bilidades da seqüência Y. A medida que aumenta o tamanho da janela aumenta também o valor das probabilidades encontradas. Isso é visto novamente no gráfico da figura 6.10, onde são exibidas as médias (linha contínua dos gráficos) das probabilidades médias dos valores da tabela $\{\text{scrDEF}\}$ em todos os arquivos de escore para todas as seqüências de classificação nos vários tamanhos de janela considerados. A maior média foi 0,857 para uma janela de tamanho 129.

Os tempos gastos para classificar as seqüências podem ser vistos na tabela 6.8. O tempo médio encontrado para classificar uma seqüência em cada tamanho de k é exibido na segunda coluna e o tempo total gasto para classificar todas as seqüências em cada tamanho de k é mostrado na terceira coluna. O grande aumento no tempo médio para $k = 129$ se deve a enorme quantidade de eventos considerados no calculo dos valores das probabilidades de cada base de uma seqüência.

k	tempo médio(s)	tempo total(s)
1	0,136698	26,246000
3	0,170068	32,653000
5	0,227328	43,646996
9	0,424599	81,523003
17	1,122578	215,534897
33	3,717333	713,727905
65	13,620148	2615,068359
129	52,058208	9995,175781

Tabela 6.8: Tempo de execução do programa `Pre` para classificar todas as seqüências dos conjuntos 4 a 6.

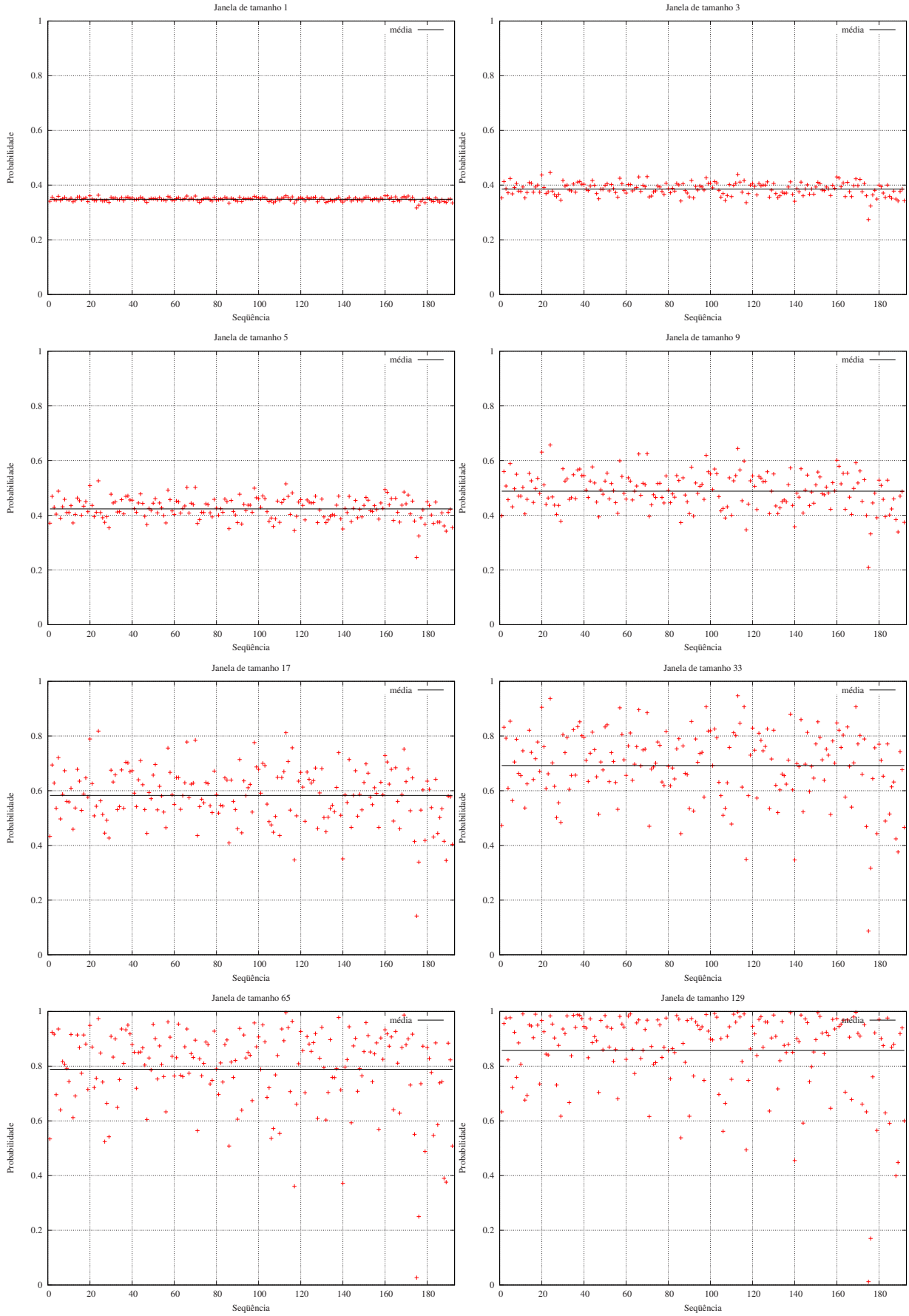

Figura 6.9: Médias das probabilidades de todos os campos da tabela {scrDEF} para cada uma das janelas em todos os arquivos.

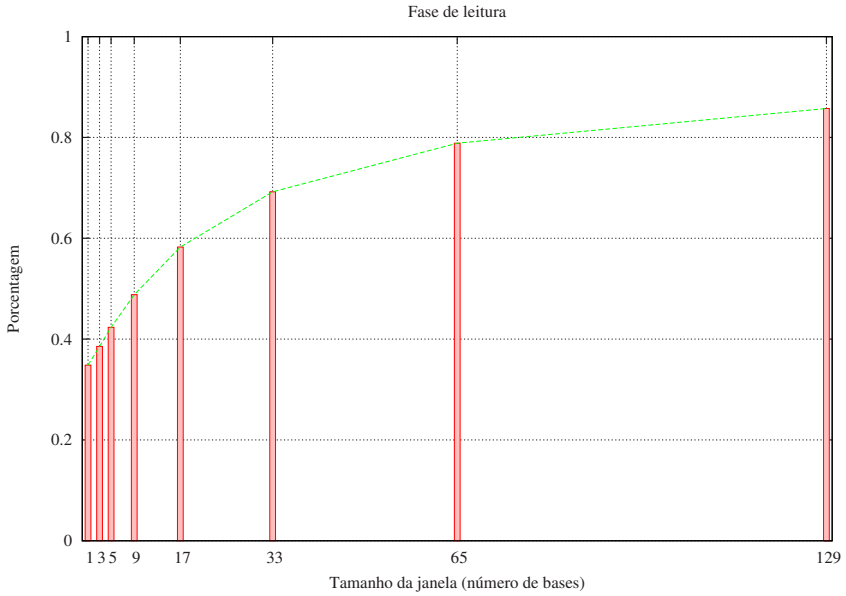

Figura 6.10: Identificação da correta fase de leitura no genoma humano.

Capítulo 7

Conclusões e trabalhos futuros

Abordagem Neste trabalho propomos um método computacional baseado em características estatísticas para resolver o problema da identificação das regiões codificadoras em seqüências de DNA.

O método desenvolvido emprega o Teorema de Bayes. As probabilidades necessárias para a aplicação do teorema (*a priori* e condicionais) são obtidas através da decomposição em seqüências de comprimentos iguais a 1, 2 e 3 bases, dispensando o uso de tabelas muito grandes cuja construção exigiria uma excessiva quantidade de amostras com classificação previamente conhecida.

Por sua natureza, o método detecta tanto regiões codificadoras ativas quanto regiões codificadoras fósseis (que foram ativas em organismos ancestrais) nos trechos não-codificadores. Por esta característica, ele pode ajudar na dedução do estudo da recuperação da informação da história dos genomas analisados.

O desempenho do método pode ser ajustado variando-se o tamanho k da janela usada. Uma janela pequena reduz o custo de processamento e permite a detecção de trechos codificadores menores; por outro lado a classificação obtida tem baixa confiabilidade.

O método fornece uma classificação probabilística das bases de uma cadeia de DNA em eventos do tipo: N (base não-codificadora), D (base do início de um códon), E (base do meio de um códon) e F (base do final de um códon).

Para uma análise mais precisa das potencialidades do método apresentado seria desejável uma maior quantidade de testes com bases de dados genômicas de outros organismos.

Desempenho Resultados significativos foram alcançados mostrando a viabilidade do método: as bases codificadoras no genoma humano possuem em média uma probabilidade de 75% de serem rotuladas como tal e cada base dentro do códon recebe uma probabilidade média de 85% de que seu rótulo seja o correspondente a sua fase correta. Porém, resultados melhores poderiam ser encontrados com a obtenção de amostras para o treinamento com uma classificação mais precisa, com estudos dos indícios das localizações das regiões codificadoras fosséis dentro das regiões não-codificadoras.

A performance poderia ser aumentada com a incorporação de características biológicas já conhecidas e que são utilizadas pelas próprias estruturas celulares no reconhecimento das regiões codificadoras, tais como: localização dos códons de parada e de início, localização dos promotores etc.

Trabalhos futuros Nosso método tem dificuldade em determinar precisamente fronteiras entre as regiões codificadoras e não-codificadoras. Verifica-se entretanto que as fronteiras íntron-éxon tem características especiais. Em particular há uma alta freqüência dos pares GT e AG [15] imediatamente antes e depois de cada éxon. Estas características poderiam ser levadas em conta, por exemplo introduzindo rótulos especiais para estes marcadores.

Além disso, poderíamos estender o modelo de DNA utilizado para considerar janelas com mais de uma quebra diminuindo o tamanho mínimo das regiões codificadoras e não-codificadoras, permitindo a identificação mais precisa em regiões curtas com janelas maiores.

Visto que o modelo aqui proposto para as regiões não-codificadoras não representa adequadamente suas características estatísticas, podemos considerar como uma grande melhora a incorporação de um modelo mais realista para estas regiões.

Referências

[1] Banco de dados com as seqüências do homem. http://www.fruitfly.org/seq_tools/datasets/Human/multi_exon_GB.dat.

[2] Genbank - banco de dados de genomas. http://www.ncbi.nlm.nih.gov/Genbank/.

[3] Genbank - estatísticas do crescimento de seqüências armazenadas. http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html.

[4] ALBERTS, B., BRAY, D., LEWIS, J., RAFF, M., ROBERTS, K., WATSON, J. D. *Biologia Molecular da Célula*, 3 ed. Editora Artes Médicas Sul, 1997.

[5] BATZOGLOU, S., PACHTER, L., MESIROV, J. P., BERGER, B., LANDER, E. S. Human and mouse gene structure: Comparative analysis and application to exon prediction. *Genome Research*, 10 (2000), 950–958.

[6] COOPER, G. M. The cell: A molecular approach. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Search&db=books&doptcmdl=GenBookHL&term=chromosome+AND+cooper%5Bbook%5D+AND+165269%5Buid%5D&rid=cooper.table.619, 2000.

[7] FICKETT, J. W. Recognition of protein coding regions in DNA sequences. *Nucleic Acids Research 10*, 17 (1982), 5303–5318.

[8] GISH, W., STATES, D. J. Identification of protein coding regions by database similarity search. *Nature Genetics*, 3 (1993), 266–272.

[9] GRIBSKOV, M., DEVEREUX, J., R. BURGESS, R. The codon preference plot: Analysis of protein coding sequences and prediction of gene expression. *Nucleic Acids Research 12*, 1 (1984), 539–549.

[10] HUMAN GENOME SEQUENCING CONSORTIUM INTERNATIONAL. Finishing the euchromatic sequence of the human genome. http://www.nature.com/cgi-taf/DynaPage.taf?file=/nature/journal/v431/n7011/full/nature03001_fs.html, October 2004.

[11] JUNQUEIRA, L. C., CARNEIRO, J. *Biologia Celular e Molecular*, 6 ed. Guanabara koogan, 1997.

[12] KOTLAR, D., LAVNER, Y. Gene prediction by spectral rotation measure: A new method for identifying protein-coding regions. *Genome Research 13*, 8 (2003), 1930–1937.

[13] MAJOROS, W. H., PERTEA, M., SALZBERG, S. L. Efficient implementation of a generalized pair hidden Markov model for comparative gene finding. *Bioinformatics 21*, 9 (2005), 1782–1788.

[14] MEIDANIS, J., SETUBAL, J. C. *Introduction to Computational Molecular Biology*. PWS Publishing Company, 1997.

[15] NICORICI, D., ASTOLA, J., TABUS, I. Computational identification of exons in DNA with a hidden Markov model. In *GENSIPS, Workshop on genomic signal processing and statistics* (Raleigh, North Carolina, USA, October 2002).

[16] PAPOULIS, A. *Probability, Random Variables, and Stochastic Processes*, 3 ed. McGraw-Hill, 1991.

[17] PERTEA, M., SALZBERG, S. L. Computational gene finding in plants. *Plant Molecular Biology 48*, 1-2 (2002), 39–48.

[18] ROGOZIN, I. B., D'ANGELO, D., MILANESI, L. Protein-coding regions prediction combining similarity searches and conservative evolutionary properties of protein-coding sequences. *Gene 226*, 1 (January 1999), 129–137.

[19] SALZBERG, S. L. Locating protein coding regions in human DNA using a decision tree algorithm. *Journal of Computational Biology 3*, 2 (1995), 473–486.

[20] SNYDER, E. E., D.STORMO, G. Identification of coding regions in genomic DNA sequences: An application of dynamic programming and neural networks. *Nucleic Acids Research 21*, 3 (December 1993), 607–613.

[21] SNYDER, E. E., D.STORMO, G. Identification of protein coding regions in genomic DNA. *Journal of Molecular Biology*, 248 (1995), 1–18.

[22] STADEN, R., MCLACHLAN, A. D. Codon preference and its use in identifying protein coding regions in long DNA sequences. *Nucleic Acids Research 10*, 1 (1982), 141–155.

[23] WANG, Y., ZHANG, C.-T., DONG, P. Recognizing shorter coding regions of human genes based on the statistics of stop codons. *Biopolymers 63*, 3 (2002), 207–216.

[24] WU, Y., LIEW, A. W.-C., YAN, H., YANG, M. Classification of short human exons and introns based on statistical features. *Physical Review E*, 67 (2003).