UNIVERSIDADE ESTADUAL DE SANTA CRUZ

PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E BIOLOGIA MOLECULAR

IDENTIFICAÇÃO COMPUTACIONAL DE GENES INDEPENDENTE DE ESPÉCIE EM GENOMAS EUCARIOTOS

RAMON OLIVEIRA VIDAL

ILHÉUS – BAHIA – BRASIL
Setembro de 2006

RAMON OLIVEIRA VIDAL

IDENTIFICAÇÃO COMPUTACIONAL DE GENES INDEPENDENTE DE ESPÉCIE EM GENOMAS EUCARIOTOS

Dissertação apresentada à Universidade Estadual de Santa Cruz, como parte das exigências para obtenção do título de Mestre em Genética e Biologia Molecular.

Área de concentração: Biotecnologia e Genômica.

ILHÉUS – BAHIA – BRASIL
Setembro de 2006

V649 Vidal, Ramon Oliveira.

Identificação computacional de genes independente de espécie em genomas eucariotos / Ramon Oliveira Vidal. – Ilhéus, Ba : UESC, 2006.

x, 69f. : il.

Orientador: Diego Gervasio Frias Suárez.

Dissertação (Mestrado) – Universidade Estadual de Santa Cruz. Programa de Pós-graduação em Genética e Biologia Molecular.

Bibliografia: f. 48-53.

Inclui apêndices.

1. Biotecnologia. 2. Genética. 3. Genoma. 4. Biodiversidade. I. Título.

CDD 660.6

RAMON OLIVEIRA VIDAL

IDENTIFICAÇÃO COMPUTACIONAL DE GENES INDEPENDENTE DE ESPÉCIE EM GENOMAS EUCARIOTOS

Dissertação apresentada à Universidade Estadual de Santa Cruz, como parte das exigências para obtenção do título de Mestre em Genética e Biologia Molecular.

Área de concentração: Biotecnologia e Genômica.

APROVADA: 29 de setembro de 2006

Prof Dr. Dario Ahnert
UESC/DCB

Prof. Dr. Ronan Xavier Corrêa
UESC/DCB

Prof. Dr. Rubem Pinto Mondaini
UFRJ/COPPE

Prof. Dr. Diego Gervasio Frías Suárez
UESC/DCET – Orientador

DEDICATÓRIA

Aos meus pais que, com muito apoio e com todas as dificuldades, confiaram em mim, dedicando grande parte de suas vidas em função de meu crescimento social, intelectual e espiritual. Sem a educação familiar eu não chegaria tão longe.

AGRADECIMENTOS

A Deus.

À Universidade Estadual de Santa Cruz, pela oportunidade da realização do curso de graduação e mestrado.

Ao CNPq e FAPESB que me deram apoio num período que tanto precisei.

Ao Professor Diego Frias que me deu oportunidade de iniciar na pesquisa cientifica, acreditando em meu potencial. Agradeço também pela orientação, pela amizade e pelo apoio recebido em todos esses anos na UESC.

Ao Prof. Júlio Cascardo, pela orientação e amizade e apoio recebido, principalmente nas áreas biológicas; seus elogios sempre foram incentivadores do meu trabalho.

Ao Dr. Nicolas Carels, pela amizade e orientação recebidas durante todo o período que esteve na UESC como pesquisador visitante. É uma pessoa de extrema competência e foi fundamental para que pudesse produzir tantos bons resultados.

Aos grandes amigos Bruno Leite, Matheus Thomaz e Valdemir Junior pela amizade, que apesar de distantes ainda me incentivam e me apóiam.

Aos Colegas do mestrado e do laboratório, pela amizade e pelo apoio.

Às secretárias do Mestrado e PROPP, Luciana e Alzira, pelo suporte em momentos de ansiosidade.

À Ana Cácia, pelo apoio, amizade, companheirismo e carinho.

A todas as pessoas que me apoiaram e incentivaram direta ou indiretamente durante esse período inicial de minha vida acadêmica.

EXTRATO

VIDAL, Ramon Oliveira, Universidade Estadual de Santa Cruz, Ilhéus, setembro de 2006. **Identificação computacional de genes independente de espécie em genomas de eucariotos.** Orientador: Diego Gervasio Frías Suárez. Co-orientador: Julio César de Matos Cascardo. Co-orientador: Nicolas Carels.

O presente trabalho teve como objetivo inicial o estudo, a avaliação e a adaptação de métodos computacionais para detecção de genes *ab initio* que não requerem informações sobre o organismo em estudo nem o conhecimento prévio da espécie, tendo baixa sensibilidade ao conteúdo de GC e ao uso do códon. Numa segunda etapa objetivou-se o desenvolvimento de ferramentas computacionais para a identificação das regiões codificantes de qualquer espécie a serem utilizadas em larga escala no Laboratório de Bioinformática da UESC (LABBI) em projetos de anotação genômica e ESTs, tanto de novos organismos como de misturas de seqüências de organismos desconhecidos, como é o caso do metagenoma.

Foi escolhido um método de baixo custo computacional capaz de detectar a organização do DNA codificante em códons, em particular, o método da *Average Mutual Information* (AMI) baseado na entropia de Shannon da teoria da informação. É uma medida do poder codificante de uma seqüência de DNA e permitem resolver o problema de classificação entre seqüências codificantes ou não-codificantes com eficiência superior a 95% para seqüências com comprimento maior de 400 bases. A AMI utiliza-se de contagens das vezes que os nucleotídeos A, C, G, T são encontrados nas diferentes posições da seqüência de DNA. Devido a isto, o poder discriminador de técnicas de medição do potencial codificante baseadas na quantidade de entropia decai de maneira acentuada para seqüências menores que 400 bases de pares como conseqüência da perda de consistência estatística na contagem.

Uma parte importante do estudo foi na investigação do efeito da variabilidade genética induzida por mutações pontuais no poder codificante dos genes. Simulando computacionalmente um processo acumulativo de mutação aleatória das bases de seqüências codificantes, foi observado que o poder codificante decai rapidamente quando a taxa de mutação cresce, chegando ao nível das seqüências randômicas quando 30% dos nucleotídeos sofreram mutações, o que coincide com o estágio em que aproximadamente 15% dos codons das seqüências tiveram duas das suas três bases mutadas. Entender o efeito das mutações na evolução dos genes é importante no desenvolvimento de ferramentas computacionais para análise de seqüências qualquer que seja a metodologia abordada.

Esses métodos independentes da espécie não permitem determinar o quadro de leitura das seqüências que são classificadas como codificantes. Para solucionar esta limitação foi desenvolvido um novo método derivado da própria teoria da

informação mútua aplicando um procedimento de perturbação do sinal de entrada que conduziu à descoberta de novas variáveis mais sensíveis à estrutura de codons do que a própria AMI. O novo método, chamado de FRAMER, procura a posição, dentre as três possíveis, onde a quantidade de informação é maior, identificando corretamente o quadro de leitura. O FRAMER é um método robusto dado que: (i) é aplicável diretamente para qualquer espécie; (ii) atinge taxas de acerto da ordem de 90% inclusive com seqüências curtas de 50 pares de bases e (iii) não é afetado pela presença de regiões não codificantes, por exemplo, introns, sempre que a região codificante não seja inferior à quarta parte do comprimento total da seqüência analisada. Estas características foram confirmadas no processamento de ESTs do arroz, que são seqüências curtas de 200 bp de comprimento médio, contaminadas frequentemente com regiões não codificantes pertencentes ao vetor de clonagem e com elevadas taxas de erros de sequenciamento que dificultam o uso de outras técnicas convencionais.

As novas ferramentas de bioinformática desenvolvidas colocam à Universidade Estadual de Santa Cruz em uma vantajosa condição, fornecendo o suporte computacional necessário para o processamento da informação proveniente de projetos de sequenciamento genômico e de ESTs de novas espécies, abrindo as portas para o estudo sistemático e em larga escala da biodiversidade regional.

Palavras-chave: identificação de genes, anotação de genomas, informação mútua, quadro de leitura, biodiversidade.

ABSTRACT

VIDAL, Ramon Oliveira, Universidade Estadual de Santa Cruz, Ilhéus, setembro de 2006. **Computational Species Independent Gene Finding In Eukaryote Genomes.** Advisor: Diego Gervasio Frías Suárez. Advisor Committee Members: Julio César de Matos Cascardo and Nicolas Carels.

The present work had as initial objective the study, evaluation and adaptation of computational methods for *ab initio* gene detection, i.e., that do not require previous information about the studied organism such as GC content and/or codon usage. In a second stage, it was focused on the development of high throughput computational tools for the identification coding regions in eukaryotic species. These tools are needed for sequences annotation in genomics and transcriptomics (ESTs) of new organisms and organism mixtures (metagenomics) at the Bioinformatics Laboratory of UESC (LABBI),

For this study, we used Average Mutual Information (AMI) that is based on the Shannon entropy (information theory). This method is able to detect the periodicity of coding DNA. AMI allow the classification of a sequence in coding or not coding with efficiency higher than 95% provided that sequences are larger than 400 bases. AMI uses the frequency of the four nucleotides A, C, G and T, in the DNA sequence. Because of this, the discriminating power for sequences below 400 base pairs decreases quickly as a consequence of the lost of statistical consistency.

By investigating the effect of the accumulation of point mutations on the coding potential of genes we found that the coding power decreases rapidly with the increase of mutation rate. With 30% of mutated nucleotides, the classification between coding sequences and non-coding sequences already becomes impossible. With this mutation level, we found that 15% of codons were altered in two of their nucleotides. Understand the mutation effect in the gene evolution is important in the development of computational toold for DNA analysis.

The species independent methods do not allow the determination of the reading frame of the coding sequences. To solve such a limitation a new method derived from the mutual information theory was developed. This was obtained by the introduction of perturbations in the sequence and by studying the AMI responses. The new derived method was called FRAMER, it detects the position among three consecutive nucleotides that is associated with the highest amount of information. This position indicates the reading frame.The FRAMER is robust in the sense that: it (i) can be used for any species; (ii) reaches high efficiency, in the order of 90%, even with sequences as short as 50 bp and (iii) is not affected by the presence of non-coding regions provided that their account for more than 2 times the effective coding sequence. These characteristics had been shown with EST sequences of rice. ESTs are short sequences of 300 bp of average size, frequently contaminated with non-coding regions from due to random sequencing. This property makes the use of conventional techniques difficult.

The new developed bioinformatics tools supply the computational facilities of University State of Santa Cruz for the processing of the information resulting from genomic and transcriptomic projects of new species. They will allow the systematic and high throughput study of the regional biodiversity.

Keywords: gene finding, genome annotation, mutual information, reading frame, biodiversity

ÍNDICE

1. INTRODUÇÃO

Desde o inicio do Projeto Genoma Humano em 1990, os bancos de dados de seqüências do Homem e Organismos Modelo tem crescido muito rapidamente. Devido a essa grande quantidade de informação muitos sistemas para predição computacional de genes têm sido desenvolvidos para análise automática e anotação genômica.

A UESC, a partir do desenvolvimento do projeto genoma do *Moniliophtera perniciosa* (*Crinipellis perniciosa*), se tornou um centro de estudos de genomas regionais explorando a biodiversidade local e buscando soluções moleculares para o controle a pragas como a "*vassoura de bruxa"* causada por esse fungo (SCARPARI et al, 2005). A bioinformática necessária para a análise desses organismos deve utilizar, então, abordagens que possibilitem lidar com genomas desconhecidos, com pouca informação prévia (uso do códon, conteúdo de GC, entre outros) e provavelmente com famílias gênicas ainda não existentes nos bancos de dados públicos mundiais.

Um dos principais desafios da bioinformática é a identificação dos genes responsáveis pela codificação das proteínas, estas realizam as mais variadas funções nos organismos, desde o transporte de nutrientes e metabólitos à catálise de reações biológicas (GRIFFITH et al, 2004, p. 275), possuindo grande importância científica e econômica. Os genes podem ser divididos em duas regiões principais, a seqüência que será transcrita em um RNA mensageiro (mRNA) e a seqüência que regula a expressão desse gene, chamado também de região promotora. O mRNA poderá ser formado por exons e íntrons (MATTICK, 1994), estes últimos estão presentes principalmente em eucariotos, sendo muito raros em procariotos (MARTINEZ-ABARCA; TORO, 2000) e geralmente são extraídos do mRNA num processamento pós-transcricional chamado de *splicing*, então, a região a ser codificada pelo ribossomo será apenas a junção dos exons (BURATTI; BARALLE, 2004).

Devido ao processo da evolução biológica, onde se supõe que todas as espécies existentes atualmente (e as que já se extinguiram) têm um ancestral comum, existe uma alta similaridade entre os genomas dos distintos organismos, sendo os diferentes caminhos evolutivos os responsáveis pela diferenciação. Sendo assim, muitas ferramentas computacionais levam em consideração características específicas dos genomas das famílias taxonômicas para identificar e analisar os genes (LEWIS et al, 2000).

O conhecimento prévio da preferência do uso de códon para codificar os 20 aminoácidos existentes torna a estrutura gênica previsível em cada espécie. Ferramentas probabilísticas podem detectar genes a partir de um conjunto de treinamento ou uma matriz de dados derivados do uso do códon do organismo em estudo (SALZBERG et al, 1999) (SAUL; BATTISTUTTA, 1988). Entre outras características, a organização das seqüências codificantes em códons origina uma periodicidade, de período “três”, nas funções de correlação espacial entre nucleotídeos, que permite também a detecção dos genes (TIWARY et al, 1997) (GROSSE et al, 2000b). Essa abordagem é particularmente útil para o estudo do genoma do *Moniliophtera perniciosa*, dada a dificuldade em se obter uma amostra representativa de genes validados para serem utilizados como conjunto de treinamento das ferramentas probabilísticas.

Genes com seqüências nucleotídicas similares traduzem proteínas de funções também similares, essa conservação permite identificar genes homólogos em genomas distintos inferindo consequentemente sua função (ALTSCHUL et al, 1990). Cerca de 50% dos genes de um organismo pode ser identificado dessa forma, mas esse tipo de abordagem somente identifica as regiões mais conservadas do gene que devido a sua importância na proteína não podem sofrer mutações. As regiões iniciais e finais do gene, assim como os limites dos exons são pouco conservados, sendo a predição dessas regiões pouco eficazes com esse método (MATHE et al, 2002).

Identificar as regiões de fronteira entre íntrons e exons, regiões promotoras, códons iniciais e terminais, além de delimitar o gene pode ser mais uma confirmação dele e, no caso dos promotores, caracterizarem a expressão da proteína. Para uma identificação geral dos promotores o mais indicado para ser independente da espécie é buscar somente o promotor basal por ser conservado e estar presente na maioria dos genes eucariotos (FICKETT; HATZIGEORGIOU, 1997; KNUDSEN, 1999). Outro motivo secundário é que no início dos projetos genomas os *contigs* são muito pequenos para localizarmos regiões do promotor que podem ser muito distantes da região inicial da transcrição.

Esta dissertação teve como objetivo desenvolver ferramentas para identificação de genes independente da espécie para aplicação no processo de anotação de novos genomas. A primeira parte contém uma revisão bibliográfica sobre identificação computacional de genes. Após isso, a dissertação foi dividida em dois capítulos. O primeiro é uma análise do método de identificação de regiões codificantes independente de espécie chamado de *Average Mutual Information* (GROSSE et al, 2000a) e uma avaliação do impacto das mutações na quantidade de informação codificante. O segundo capítulo mostra o desenvolvimento de uma ferramenta para detecção do quadro de leitura independente da espécie. Essa dissertação possui ainda dois apêndices mostrando, no primeiro, uma aplicação da *Mutual Information* na certificação de genes do arroz para filtrar os falsos positivos gerados pelo processo de anotação automática com ferramentas de aprendizagem e, no segundo, o desenvolvimento de uma ferramenta baseada em rede neural multicamadas perceptron para detecção do promotor basal utilizando o padrão do *TATA-box* existente nos promotores eucariotos, essa ferramenta visa dar mais subsídios na detecção da região codificante.

2. REVISÃO DE LITERATURA

2.1 Características dos genomas e Identificação computacional de genes

Genes que codificam proteínas são os mais estudados nos projetos genoma. Estes são responsáveis pelas proteínas expressas de um organismo, possuindo assim, uma grande importância científica e econômica. Essas seqüências gênicas possuem uma região com informações da seqüência de aminoácidos da proteína que será transcrita e uma região reguladora, ou promotor (Fig. 1), que indica onde, quando e em que intensidade esse gene será transcrito e expresso.

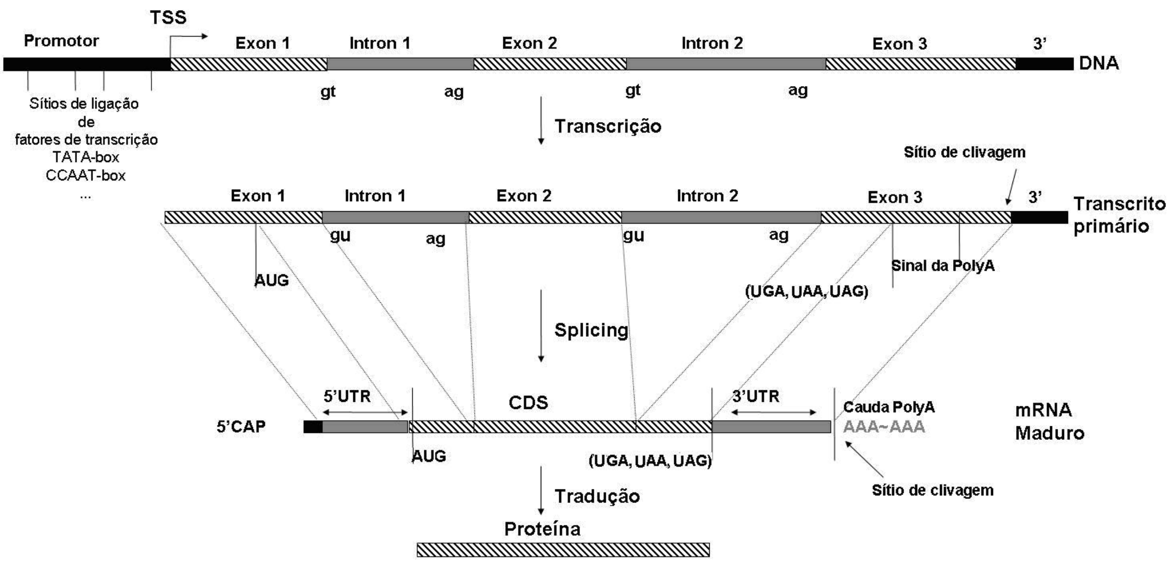


Figura 1 – Estrutura da transcrição e tradução em genes de eucariotos. Na primeira linha esta a representação do gene no DNA genômico, com os íntrons e éxons e a região promotora. A próxima linha representa o mRNA primário com os íntrons ainda presentes, nota-se que as timinas (T) são substituídas por uracilas (U). A linha abaixo é uma representação do mRNA maduro após ter sofrido o processo de *splicing* (remoção dos íntrons), acréscimo da cauda polyA na extremidade 3´ e da proteína CAP na extremidade 5`. A última linha é a representação da proteína, após o mRNA maduro ter sido traduzido pelo ribossomo.
Fonte: http://www.nslij-genetics.org/gene, com modificações.

Genes procariotos e eucariotos seguem o mesmo código (Tabela 1) que é degenerado, ou seja, mais de um códon pode codificar um mesmo aminoácido, gerando repetições nas seqüências gênicas. Apesar dessa semelhança os genes eucariotos e procariotos se diferenciam em vários aspectos que devem ser levados em conta no desenvolvimento de ferramentas computacionais: (i) os genomas procariotos são mais condensados, os genes estão mais próximos podendo haver sobreposições de genes com maior freqüência (MATHE et al., 2002), (ii) os promotores procariotos são mais conservados, menos complexos e o mesmo promotor pode regular vários genes na estrutura conhecida como operon, que ocorre em raras exceções em eucariotos (BLUMENTHAL, 2004) e (iii) organização dos genes eucariotos em introns e exons (Figura 1), sendo que após a transcrição os íntrons são clivados e os éxons são unidos, resultando no mRNA maduro.

Tabela 1 – Código genético

	U	C	A	G	
U	UUU — Phe UUC — Phe UUA — Leu UUG — Leu	UCU — Ser UCC — Ser UCA — Ser UCG — Ser	UAU — Tyr UAC — Tyr UAA — **stop** UAG — **stop**	UGU — Cys UGC — Cys UGA — **stop** UGG — Trp	U C A G
C	CUU — Leu CUC — Leu CUA — Leu CUG — Leu	CCU — Pro CCC — Pro CCA — Pro CCG — Pro	CAU — His CAC — His CAA — Gln CAG — Gln	CGU — Arg CGC — Arg CGA — Arg CGG — Arg	U C A G
A	AUU — Ile AUC — Ile AUA — Ile AUG — Met	ACU — Thr ACC — Thr ACA — Thr ACG — Thr	AAU — Asn AAC — Asn AAA — Lys AAG — Lys	AGU — Ser AGC — Ser AGA — Arg AGG — Arg	U C A G
G	GUU — Val GUC — Val GUA — Val GUG — Val	GCU — Ala GCC — Ala GCA — Ala GCG — Ala	GAU — Asp GAC — Asp GAA — Glu GAG — Glu	GGU — Gly GGC — Gly GGA — Gly GGG — Gly	U C A G

Esta característica de processamento pós-transcricional em eucariotos torna a busca por ORFs (*open reading frame*) ineficiente, já que um intron pode conter códons de parada (*stop codons*) ou deslocar o quadro de leitura do próximo exon. Em eucariotos mais complexos podem ocorrer clivagens alternativas, o que confere a esses organismos uma maior variedade de produtos protéicos mesmo possuindo poucos genes (ZHANG, 2002). A quantidade, disposição e tamanho dos introns por gene variam entre os organismos e de um gene para outro (VINAGRADOV, 2001).

Para tratar dessa complexidade na estrutura gênica, existem várias abordagens para identificação computacional de genes que podem ser agrupadas em dois princípios básicos: (i) identificação da região codificante de proteína, que é a classificação de uma seqüência como codificante e não-codificante, também conhecidos como sensores de conteúdo e (ii) identificação de sítios funcionais do gene, como o primeiro códon do gene, promotores, a região de ligação da cauda poliA e regiões de *splicing* (MATHE at al, 2002).

Dentro dessas duas grandes abordagens existem os métodos baseados em similaridade, aprendizado de máquina e as técnicas *ab initio* que são independentes de qualquer informação prévia e alinhamento. Esta última é de grande importância no estudo de novos genomas, tanto na fase inicial quando não existem muitos dados, quanto na fase final para certificar os resultados.

Descobrir novos genes, suas funções, incluindo regiões relevantes para a sua expressão como promotores e sítios de splicing é uma tarefa que requer a utilização de abordagens integradas, nenhum conjunto de software existente inclui todas as ferramentas necessárias, e cada uma tem suas limitações (FICKETT, 1996; MATHE et al, 2002).

2.2. Abordagens computacionais para identificação de genes

2.2.1 Abordagem por similaridade ou extrínseca

Essa abordagem é baseada no princípio de que regiões funcionais (éxons) são mais conservadas evolutivamente que regiões não-funcionais (regiões intergênicas e intrônicas).

Sua principal vantagem é que já há informações biológicas sobre o gene predito. No entanto, apresenta as desvantagens de ter a consistência de seus resultados dependente da quantidade e qualidade do banco de dados utilizado na comparação e não ser capaz de descobrir novas famílias de genes, identificando somente as regiões mais conservadas entre as espécies (CARELS et al, 2004). Os limites do gene que são mais sujeitos a modificação também não são preditos com tanta eficácia.

Gelfand et al., 1996 foi o pioneiro na identificação de genes por similaridade com o software *Procrustes*. Após o *Procrustes* muito outros foram desenvolvidos e todos eles são sofisticações do algoritmo de alinhamento local tradicional de *Smith-Waterman* (SMITH et al, 1981). O mais famoso e utilizado para detecção de genes por similaridade é o BLAST (ALTSCHUL et al, 1990), principalmente pela sua alta velocidade em fazer alinhamentos e baixo custo computacional.

Aproximadamente metade dos genes pode ser predito por similaridade, esse valor tende a crescer conforme novos projetos genoma sejam concluídos e novos genes depositados nos banco de dados públicos (ZHANG et al, 2002).

2.2.2 Sensores de conteúdo intrínseco dependente da espécie

São métodos que utilizam de conhecimentos prévios da espécie analisada, geralmente o uso do códon, para poder encontrar novos genes. Esse conhecimento pode ser obtido através de uma ferramenta de aprendizado, sendo necessário ter uma amostra suficientemente representativa de genes confirmados, de preferência experimentalmente, o que não é possível em novos genomas de espécies pouco estudadas.

Metodologias baseadas no uso do códon leva em conta a característica em que os organismos não possuem todos os tRNAs em igual proporção para codificar os 20 aminoácidos, eles optam por um conjunto específico de RNAs transportadores que também é correlacionado com o teor de GC do genoma (Kanaya et al, 2001). Esta característica confere uma regularidade à região codificante gerando padrões que podem ser explorados de diversas maneiras.

Contudo, um pequeno erro na fase de captura do padrão ou de aprendizado pode resultar em predições erradas (CRUVEILLER et al., 2004; CARELS et al., 2004). A quantidade e qualidade dos dados utilizados nessa fase também podem resultar numa perda da capacidade de generalização ou numa super-especialização (ABU-HANNA, 1999).

As ferramentas mais utilizadas nesse tipo de abordagem são os modelos ocultos e interpolados de Markov e as redes neurais artificiais.

As cadeias de Markov são conjuntos de métodos para computar a probabilidade de um evento baseado em número fixo de eventos prévios. No contexto da análise de seqüências de DNA, as cadeias de Markov prevêem uma base, examinando um número fixo de bases prévias àquela base na seqüência. Por exemplo, uma cadeia de Markov de ordem cinco irá prever uma base examinando as cinco outras prévias, estratégia muito utilizada em genomas de bactérias (BORODOVSKY and MCININCH, 1993; BESEMER et al., 2005). Modelos Ocultos de Markov são chamados assim quando a seqüência de estados não é observada, ao invés disso, as saídas geradas pelos estados é que são observadas. Em seqüências genômicas, por exemplo, esses estados podem ser os éxons ou íntrons. As saídas geradas por cada estado dependerão do estado atual e também das saídas geradas por estados prévios. No caso da busca por genes, as saídas de cada estado é uma das quatro bases A, T, C ou G. A probabilidade de cada base dependerá do tipo de estado (Guigó et al, 2000).

O Modelo Interpolado de Markov (IMM) é uma generalização das cadeias de Markov de ordem fixa. A principal diferença é que esses modelos utilizam números variáveis de bases para cada predição, em alguns contextos utilizará as 5 bases, enquanto em outros contextos poderá utilizar seis ou mais bases, ou 4 ou menos bases. O IMM busca qual oligômero ocorre com mais freqüência para fazer predições mais eficientes, ele tem a capacidade de combinar cadeias de Markov de varias ordens. IMMs são a base do GLIMMER e GlimmerM, para procariotos e eucariotos respectivamente (SALZBERB et al, 1999) (DELCHER et al., 1999).

As redes neurais artificiais, assim como as cadeiras de Markov, são ferramentas de múltiplo uso e podem ser utilizadas em vários tipos de abordagens, e inclusive como integradoras delas. Redes neurais podem utilizar um conjunto de treinamento de genes de um determinado organismo para aprender o padrão uso do códon e ser capaz de encontrar novos genes que possuem o mesmo padrão. Assim foi desenvolvido o GRAIL (UBERBACHER; MURAL, 1991) que teve três versões, a primeira identificava o potencial codificante de uma determinada seqüência, a segunda utilizava uma janela deslizante fixa e buscava informações das regiões 5’ e

3’ do exon buscando os seus limites, a última versão utilizava janela deslizante variável, buscando a região de início da transcrição, regiões terminadoras, sendo uma versão apropriada para o contexto genômica.

2.2.3 Sensores de conteúdo intrínseco independente da espécie

Os métodos baseados em de Markov e Redes Neurais necessitam de um conhecimento prévio da estrutura dos genes da espécie analisada e isso é limitante para novas espécies onde o conjunto de genes conhecidos é reduzido e certamente não é representativo de todos os potenciais genes (WANG et al, 2004).

Li et al (1992) mostrou as diversas correlações entre nucleotídeos em seqüências curtas e longas distâncias (1-3 bases e de 3-6 bases a 10 kb respectivamente), onde as correlações curtas estão relacionadas a estruturas gênicas enquanto as longas estão relacionadas com conformações estruturais na molécula do DNA como a formação dos nucleossomos (TRIFONOV et al, 1980).

É conhecido que seqüências em regiões que codificam proteínas em moléculas de DNA exibem um período 3 (três) causado pela estrutura dos códons envolvidos na tradução de seqüências de nucleotídeos em aminoácidos e que, apesar de bastante sutil, contribui no empacotamento do DNA (TRIFONOV et al, 1980). Esse período pode ser detectado e quantificado em regiões codificantes de proteínas através de metodologias baseadas no processamento de sinais.

Todos os métodos sensores de conteúdo independente de espécie utilizam de artifícios para detectar esse sinal causado pelo período 3. A predição se torna independente da espécie e de treinamento já que é um padrão universal sendo de interesse na anotação de novos organismos de genomas desconhecidos e onde existem poucos dados para análise.

2.2.3.1 Transformada de Fourier

Essa periodicidade em seqüências codificantes foi processada por Tiwary et al (1997) utilizando uma Transformada de Fourier Discreta (DFT) e identificada através do um pico na freqüência 1/3 (Figura 2 - a).

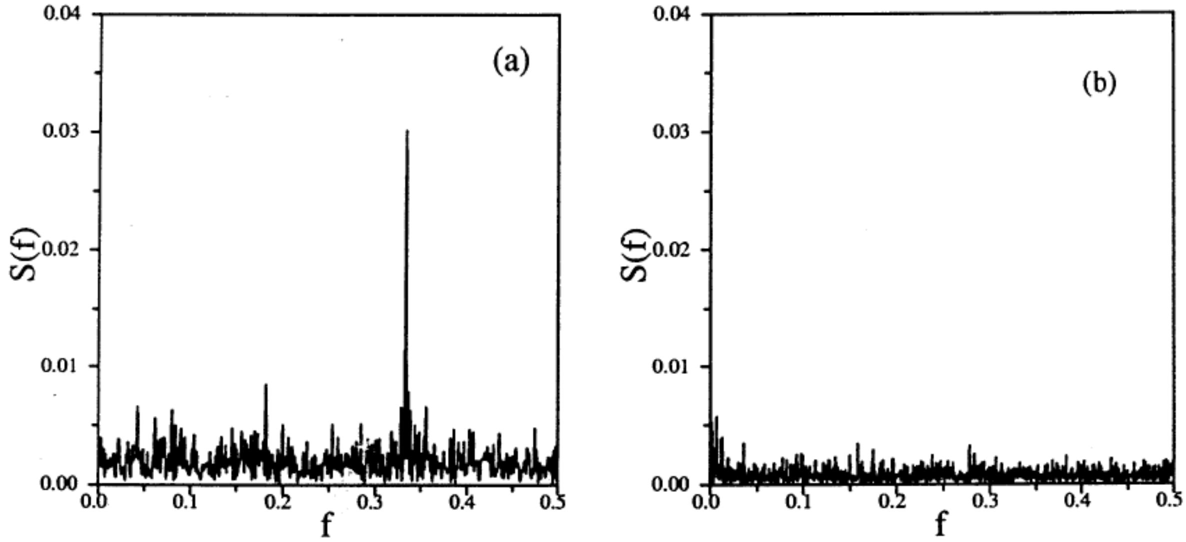


Figura 2. Típico espectro de Fourier para uma seqüência codificante (a) e não codificante (b).
Fonte: (Tiwary et al 1997).

Para realizar o cálculo da DFT é necessário codificar a seqüência de nucleotídeos formada pelos caracteres A, T, C e G em entidades numéricas. A mais comum é a codificação em quatro seqüências binária $u_A(k)$,$u_T(k)$,$u_C(k)$ e $u_G(k)$ onde k é a posição do nucleotídeo na seqüência. $u_\alpha(k)$ = 1 se na posição k existir o nucleotídeo $\alpha = (A, T, C, G)$, caso contrário $u_\alpha(k)$ = 1, como exemplificado abaixo:

Seqüência:	A	T	G	C	C	A	A	T	G	T	G	A
u_A:	1	0	0	0	0	1	1	0	0	0	0	1
u_T:	0	1	0	0	0	0	0	1	0	1	0	0
u_C:	0	0	0	1	1	0	0	0	0	0	0	0
u_G:	0	0	1	0	0	0	0	0	1	0	1	0

Aplicando DFT nestas seqüências são produzidos quatro espectros $S_A(k)$, $S_T(k)$, $S_C(k)$ e $S_G(k)$ respectivamente. O espectro da freqüência total de uma seqüência de DNA é definido por: $S \equiv |S_A(k)|^2 + |S_T(k)|^2 + |S_C(k)|^2 + |S_G(k)|^2$ (Li et al 1994), onde o pico nas seqüências codificantes é visível (Figura 2-a) enquanto nas não codificantes está ausente (Figura 2-b). A transformada de Fourier pode ser ainda aproximada computacionalmente para buscar somente a freqüência 1/3 como mostra Vaidyanathan et al (2002), reduzindo o custo computacional.

O Genescan e outros softwares derivados da DFT (ISSAC et al, 2002) são utilizados para anotações em procariotos e eucariotos (FUKUSHIMA et al, 2002) com eficiência bastante similar aos métodos que utilizam conjunto de treinamento (LYNN et al, 2001). Porém os métodos devem ser utilizados em conjunto, reduzindo o número de falsos positivos e falsos negativos que são gerados normalmente (AGGARWAL et al, 2002).

2.2.3.2 Teoria da Informação Mútua

A função da informação mútua (MI) é uma alternativa para cálculo de correlação em seqüências simbólicas (MANSILLA et al, 2002). A informação estimada pela MI resulta da ordem espacial ou entropia (SHANNON, 1948) e freqüências relativas dos quatro nucleotídeos (HERZEL et al, 1997) que difere entre seqüências codificantes e não codificantes, uma vez que a entropia é maior em seqüências não codificantes.

A entropia associada a um nucleotídeo "i" é estimada através de $Hi=Pi*logPi$, onde Pi é a probabilidade de encontrar i no sistema. A entropia associada a um par de nucleotídeos "$i\ j$" é dada por $Hij=Pij*logPij - Pi*Pj$, onde Pij denota a probabilidade dos nucleotídeos i e j estarem no sistema.

A informação mútua para dois nucleotídeos i e j é definida por $MIij= Hij-Hi*Hj$. Li et al (1992) mostrou que, considerando todos os nucleotídeos de uma seqüência, a função da informação mútua pode ser escrita da seguinte forma: $I(d) = \sum_{i,j} P_{i,j}(d) \log_2 \left(\frac{P_{i,j}(d)}{P_i \; P_j} \right)$, onde d é a distância entre dois nucleotídeos i e j. Em seqüências com alta entropia $Pij = Pi*Pj$ e, consequentemente, $I = 0$ (Figura 3). Em seqüências codificantes, as correlações entre os nucleotídeos em códons introduzem um nível de informação maior que zero (Figura 3). Isto permite utilizar a informação mutua para classificar seqüências como codificantes ou não-codificantes (GROSSE et al, 2000b). Uma abordagem prática dessa técnica foi mostrada no Apêndice A, a anotação do genoma do arroz pôde ser revisada por essa metodologia, eliminando muitos falsos positivos.

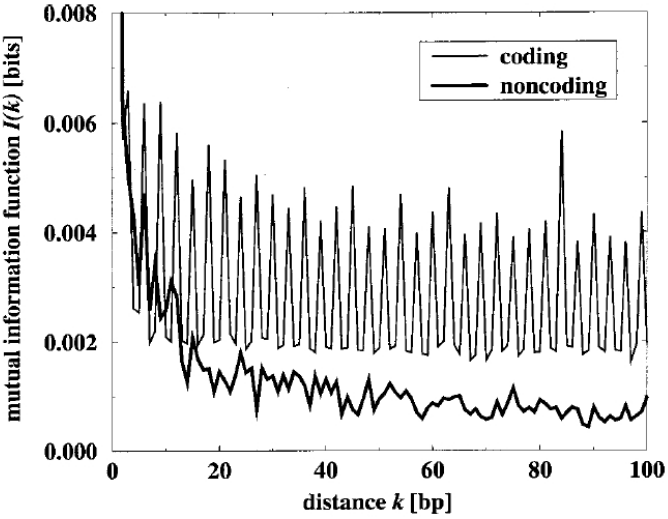


Figura 3 – Função da Informação Mútua (*mutual information*) em seqüências codificantes (linhas finas) e não codificantes (linhas grossas). Enquanto a informação ao longo da seqüência codificante oscila a cada códon numa média acima de 0, as seqüências não codificantes tendem ao valor zero. Fonte: (Grosse et al, 2000b).

2.2.3.3 Z *Curve*

O Z *curve* (ZHANG et al, 1994) é uma curva espacial tri-dimensional para visualização e análise de seqüências de DNA. Algoritmos baseados no Z *curve* têm sido aplicados para reconhecimento de seqüências codificantes no genoma humano (YAN et al, 1998), levedura (ZHANG et al, 2000), e bactérias como a Vibrio cholerae (WANG et al, 2001).

O Z *Curve* apresenta a melhor eficiência dos métodos independentes da espécie para detectar seqüências curtas, ele é capaz de amplificar a periodicidade 3 nas regiões codificantes, o Z *curve* pode aumentar a eficiência em 6-7% utilizado em conjunto com a Transformada de Fourier (YAN et al, 1998) e é simples computacionalmente.

Ele é composto de uma série de nós P0, P1, P2,..., PN, nas quais as coordenadas xn, yn e zn (n = 0, 1, 2,..., N, onde N é o tamanho da seqüência estudada), são determinadas pela chamada Z-*Transform*,

$$\begin{cases} x_n = (A_n + G_n) - (C_n + T_n), \\ y_n = (A_n + C_n) - (G_n + T_n), x_n, y_n.z_n \in [-N, N], n = 0,1,2,..., N, \\ z_n = (A_n + T_n) - (C_n + G_n) \end{cases}$$

, onde An, Cn, Gn e Tn são os números de ocorrências acumulativas de A, C, G e T, respectivamente, na subseqüência da primeira base até a n-ésima base na seqüência. Por definição A0, C0, G0 e T0 são iguais a zero para que a Z curve sempre comece pela origem. O Z curve é definido como a junção dos nós P0, P1, P2,..., PN um por um, sequencialmente, por retas (Guo et al, 2003).

Esse método introduz uma nova maneira de analisar seqüências de DNA usando uma abordagem geométrica. Além de ser a base de algoritmos para detectar genes, o Z curve permite visualizar vários padrões globais e locais do DNA e realizar comparações entre genomas (ZHANG et al, 2003).

2.2.4 Sensores de sinal

Sensores de sinais são medidas que tentam localizar sítios funcionais específicos de um gene, os principais sítios procurados são as bordas entre os introns e exons (*splicing sites*), região iniciadora e terminadora do gene e os promotores. São técnicas necessárias para refinar os resultados obtidos pelas medidas de conteúdo e que podem ser combinadas para resultados mais apurados.

Para detectar sinais é comum utilizar também de alinhamento entre a seqüência alvo e uma seqüência consenso. A seqüência consenso é determinada a partir de um alinhamento múltiplo dessas regiões em seqüências bem estudadas previamente (ROGOZIN; MILANESI, 1997; KLEFFE et al, 1996). Outra maneira de representar os sinais é através de matrizes de peso posicional que indicam a probabilidade de uma determinada base ocorrer em uma determinada posição. Essas matrizes podem ser aperfeiçoadas por redes neurais (BRUNAK et al, 1991) (REESE et al, 1997).

Para buscar possíveis características nas adjacências do sinal, como diferenças de contexto da porção do éxon e porção do íntron, modelos de Markov de maior ordem pode ser utilizados e aplicados como nos softwares VEIL (HENDERSON et al, 1997) e MORGAN (SALZBERG et al, 1998). O objetivo principal desses programas não é montar a estrutura do gene, mas indicar quais os limites dos éxons. Esses métodos podem servir para dar indicações sobre possíveis *splices* alternativos, mas ainda são pouco investigados e é um desafio futuro.

2.2.4.2 Identificação de Promotores

Identificação computacional de promotores eucarióticos é uma tarefa muito complexa, eles são caracterizados por um grande número de sítios de ligação para fatores de transcrição (TFBS; Figura 4), normalmente na região 5' do sítio de iniciação (TSS; Figura 1). Os TFBS podem ser classificados em duas estruturas: o promotor basal (*proximal promoter*) que teoricamente é comum a todos os genes e a todos os eucariotos, e um promotor *upstream* (*distal promoter*) como mostrado na Figura 4. O promotor basal fica próximo do TSS e é responsável pela captura da RNA polimerase (enzima que polimeriza o mRNA), enquanto o promotor *upstream* regula a expressão do gene de forma espaço-temporal inibindo-o ou aumentando o nível de transcrição. Esses cis-elementos são seqüências curtas (5 a 15 nucleotídeos) e degeneradas. As proteínas reguladoras (fatores de transcrição) tem afinidades com estas seqüências e se ligam no DNA para dar inicio a transcrição e regular todo o processo (FICKET; HATZIGEORGIOU, 1997; PEDERSEN et al, 1999).

O promotor basal canônico possui seqüências conhecidas por TATA-box e Inr (*Iniciator region*), enquanto o promotor upstream possui milhares de possibilidades devido a infinitas combinações dos diversos cis-elementos que podem estar também na fita complementar.

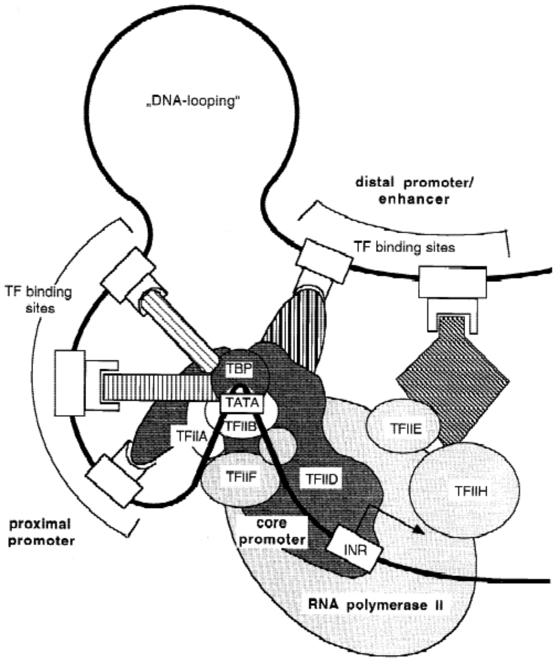


Figura 4 – Estrutura de um promotor eucariótico canônico com as proteínas da maquinaria de transcrição acopladas.
Fonte: (WERNER, 1999).

Identificação de promotores por similaridade não é muito utilizada porque as regiões mais conservadas (cis-elementos) estão inseridas em uma seqüência muito sujeita as mutações, essas regiões são muito pequenas e não estão sempre na mesma ordem, promotores de genes homólogos na maioria das vezes são diferentes (FRECH et al, 1998).

A detecção computacional da região promotora pode ser realizada buscando por (i) contextos típicos da região promotora como as ilhas CpG, presente no promotor basal de eucariotos complexos (ROMBAUTS et al, 2003), repetições internas (BELL et al, 1997) e elementos de estrutura secundária e tri-dimensional (BOHJANEN et al, 1997; KIM et al, 1995) ou (ii) buscando os cis-elementos do promotor basal ou do promotor *upstream* (WINGENDER et al, 2000) atravéz de similaridade com seqüência consenso ou com matrizes de peso posicional.

As principais ferramentas de identificação de promotores focam no reconhecimento geral do promotor, por isso utilizam apenas o promotor basal (Apêndice B) e os elementos de contexto (OHLER et al, 2001; OHLER, 2000; REESE, 2001). Os sítios de ligação a fatores de transcrição do promotor *upstream* são estudados quando se quer compreender melhor o mecanismo de regulação dos genes (FICKETT; WASSERMAN, 2000).

3. CAPITULO 1

A Influência das Mutações nas Medidas de Quantidade de Informação Codificante em Seqüências de DNA

Autores: **Ramon Vidal**, Nicolas Carels, Diego Frias.

Instituição: Laboratório de Bioinformática, Universidade Estadual de Santa Cruz Rod Ilhéus Itabuna km 16, Ilhéus, Brasil.

A Influência das Mutações nas Medidas de Quantidade de Informação Codificante em Seqüências de DNA

Ramon Vidal, Nicolas Carels, Diego Frías
Laboratório de Bioinformática, Universidade Estadual de Santa Cruz
Rod Ilhéus-Itabuna km 16, Ilhéus, Brasil
{ramon@labbi.uesc.br}

Abstract
Since the availability of high throughput techniques of DNA sequencing, the development of efficient algorithms of gene annotation became the purpose of central investigations in bioinformatics. In recent years several coding measures based on sequence entropy have being proposed for localizing coding regions in DNA sequences. In this article, we analyzed the performance of the Average Mutual Information concerning their sensibility to the threshold at which the number of random point mutations effectively destroys the codon structure detection capability. The threshold between coding and non-coding sequences represents the minimum organization level of bases in the DNA molecule that can generate a codon structure signature detectable with this kind of coding sequence measure. We found that a nucleotide substitution density of approximately 30% the measures lose their coding/non-coding discriminant values.

Introdução

Busca automática de genes é uma área de pesquisa que cresceu devido ao avanço das técnicas de seqüenciamento de DNA em larga escala. O desenvolvimento de algoritmos eficientes para detecção de genes em seqüências de nucleotídeos se tornou um problema crítico em bioinformática. Atualmente são utilizados três tipos de abordagens: (1) busca por similaridade, baseada numa busca extensiva de similaridade entre seqüências desconhecidas e seqüências conhecidas. Na maioria dos casos o BLAST (ALTSCHUL et al, 1990) é usado para este propósito. (2) Abordagens baseadas no aprendizado de máquina que implicam num processo de treinamento para aprendizado do padrão das seqüências codificantes que são específicas em cada organismo. A maioria dessas ferramentas é baseada em Modelos Ocultos de Markov (SALZBERG et al, 1999), Modelos Interpolados de Markov (KULP et al, 1996) ou Redes Neurais Artificiais (XU et al, 1996). (3) Abordagens baseadas na quantificação da entropia gerada pela organização em códons que atribuem um valor de medida de codificação para uma dada seqüência codificante (CDS). Duas medidas principais e que são correlacionadas foram propostas para identificação de regiões codificantes: a Average Mutual Information (AMI) (GROSSE et al, 2000) e a Sequence Spectral Power (SSP) (TIWARY et al, 1997).

A principal limitação da primeira abordagem é que a consistência dos resultados depende da quantidade e qualidade dos dados disponíveis em bancos públicos. Utilizando comparações com genes, esse método nunca irá descobrir novos genes com novas funções (MATHÉ et al, 2002). O desempenho da segunda abordagem é determinado principalmente pela qualidade e quantidade de dados disponíveis da espécie a ser analisada. Os classificadores devem ser treinados com dados representativos e não-redundantes para aprender os padrões das seqüências codificantes. A capacidade de generalização dessas metodologias permite a descoberta de novos genes. Infelizmente o treinamento pode produzir em muitos casos a super-especialização, ocasionando perda da capacidade de generalização (ABU-HANNA, 1999). Podem causar também o tendenciamento a algum padrão específico, ocorrendo principalmente quando a quantidade de dados para o treinamento não é representativo suficiente, levando a predição incorreta como na anotação automática do genoma do arroz (CARELS et al, 2004).

As abordagens baseadas na diferença de entropia entre seqüências codificantes e não-codificantes não dependem de banco de dados, nem de treinamento preliminar, se baseiam na característica universal da periodicidade entre os nucleotídeos dos códons. Essa característica justifica seu uso na descoberta de novos genes em seqüências de DNA (BURGE et al, 1997). Porém a presença de falsos negativos detectados por esses métodos, principalmente em organismos mais complexos, sugere que existe uma perda de potencial codificante em algumas seqüências ao longo da evolução.

Este trabalho visa quantificar, através da AMI, a perda de informação induzida por mutações pontuais simulando um modelo evolutivo. Usando software para realizar as mutações in silico, foi utilizadas seqüências de DNA de *Arabidopsis thaliana* e aplicadas substituições nucleotídicas aleatórias pontuais, simulando um evento evolutivo.

Material e Método

Um conjunto de 200 seqüências codificantes (CDS) de *Arabidopsis thaliana* experimentalmente comprovadas e um conjunto de 200 íntrons do mesmo organismo foram extraídos do Genbank. Os tamanhos das seqüências variam entre 400 e 10000 pares de bases.

A função da *Mutual Information* (SHANNON, 1948) é dada pela Equação: $M(d) = \sum\sum P_{\alpha\beta}(d)\log_2 \frac{P_{\alpha\beta}(d)}{P_\alpha\, P_\beta}$, para os símbolos $\alpha, \beta \in \{a,c,t,g\}$. $P_{\alpha\beta}(d)$ é a probabilidade que os símbolos α e β são encontrados numa distancia d um do outro.

Como discutido por Li et al, 1994, o fato de estarmos trabalhando com seqüências finitas significa que este $M(d)$ superestima $M_T(d)$ real por $M(d) - M_T(d) \approx \frac{K(K-2)}{2N}$, onde K é o numero de símbolos (quatro no caso do DNA) e N é o tamanho da seqüência.

A Average Mutual Information de uma seqüência de DNA de tamanho L como definido por Grosse et al, 2000, é uma função do poder codificante definida formalmente por: $\overline{M} = \frac{1}{D}\sum_{d=1}^{D} M(d)$, onde D, a máxima distancia analisada é um múltiplo de 3 tal que $D << L$. Na derivação de uma formula aproximada para o potencial codificante é considerado $M(3) \approx M(3n), \forall n = 1, \ldots, D/3$ e $M(4+3(n-1)) \approx M(5+3(n-1)), \forall n = 1, \ldots, D/3-1$, suportado por observações empíricas do comportamento estatístico de $M(d)$ para seqüências codificantes de varias espécies. Em uma aproximação da função da Average Mutual Information ela pode ser calculada como: $\overline{M} \approx \frac{1}{3}M(3) + \frac{2}{3}M(4)$.

Para produzir as mutações pontuais (substituições nucleotídicas), foram escolhidas e substituídas aleatoriamente nucleotídeos por outros não-redundantes em cada um dos 200 CDS da amostra. Dessa forma foram geradas 19 amostras com densidade de substituição nucleotídica crescente de 5% para 95% em acréscimos de 5%. As mutações também não foram recorrentes, ou seja, todos os níveis de mutação foram efetuados começando da seqüência original. Mutações sobrepostas foram evitadas.

Foram repetidos os experimentos três vezes para demonstrar repetibilidade nos resultados.

Todos os scripts foram programados em PERL e os gráficos feitos com o programa GRACE.

Resultados

Primeiro foi investigada a distribuição estatística da medida do poder de codificação da Average Mutual Information. A figura 1 mostra os histogramas normalizados da AMI calculada para seqüências codificantes e não-codificantes. Foi encontrada uma distancia de 1,22 inter-classe aproximadamente, que foi definida de acordo com as médias e variâncias da medida. A tabela 1 mostra as essas médias e variâncias para seqüências codificantes e não-codificantes.

Tabela 1. Médias e variâncias (entre parênteses) da AMI para seqüências codificantes e não-codificantes.

	Codificantes	***Não-codificantes***
AMI	-7.04 (1.30)	-10.84 (1.81)

Foi calculado um valor de -8.7 como discriminante da medida codificante e não codificante (Figura 1). As proporções esperadas de falsos positivos e falsos negativos para essa escolha são as mesmas, ficando satisfatória para a maioria dos propósitos práticos.

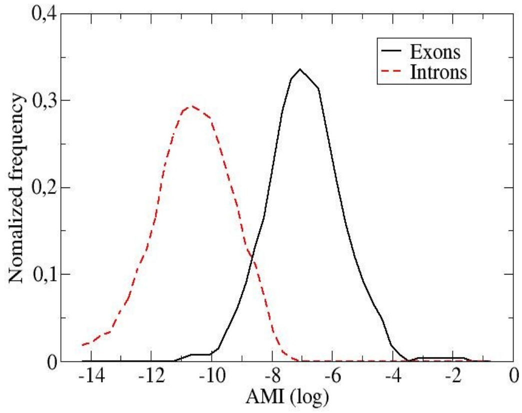


Figura 1. Distribuição da AMI para seqüências codificantes (linha contínua) e não-codificantes (tracejado).

Para explorar a aplicabilidade dessas medidas na classificação de seqüências curtas, foi feito um estudo quanto à sensibilidade ao tamanho da seqüência. A figura 4 mostra o valor médio da AMI em função do tamanho dos éxons e íntrons. A discriminação atinge a melhor eficiência a partir dos 500 bp e depois se estabiliza. Esse comportamento. Nota-se uma rápida queda no poder de discriminação a partir das seqüências com 350 nucleotídeos de comprimento.

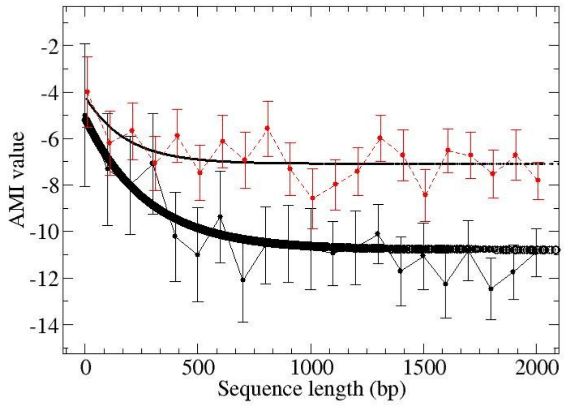


Figura 2. Valor médio da AMI para diferentes tamanhos de seqüências para seqüências codificantes (linha grossa) e não-codificante (linha fina). As barras verticais indicam os desvios padrões.

Investigando o efeito das mutações pontuais randômicas na estrutura de códons em seqüências codificantes foi encontrado um decréscimo linear na quantidade de informação das seqüências quantificada pela AMI. Todas as seqüências alcançam características de não-codificantes com aproximadamente 30% dos nucleotídeos mutados. Entretanto, nota-se que com apenas 15% de nucleotídeos mutados são produzidas medidas codificantes abaixo do valor de discriminação. A sobreposição das curvas dos três experimentos demonstra a consistência estatística dos resultados (Figura 2).

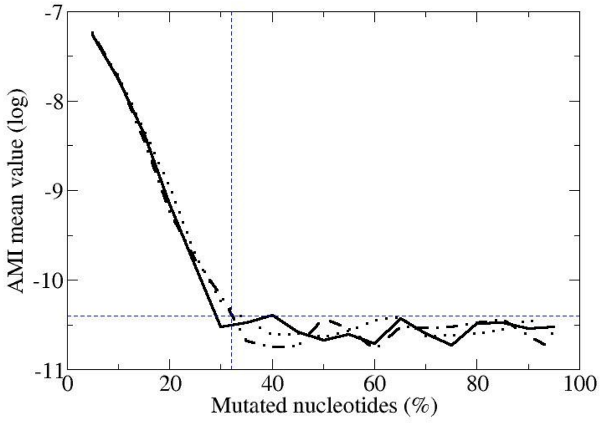


Figura 3. Distribuição da AMI para seqüências codificantes e não-codificantes. Quando a porcentagem de mutações atinge 30% a media das seqüências é a mesma das seqüências não-codificantes indicada na linha horizontal e deduzida na Fig. 1.

Observando a evolução do processo aleatório de mutações pontuais nota-se que com os 30% de mutações aleatórias, suficientes para tornar uma seqüência codificante com características de não-codificantes, cerca de 65% dos códons já tem ao menos uma substituição (Fig.3) e, dos códons mutados, 70% foram de apenas um nucleotídeo por códon, 30% duas bases substituídas e 5% apenas onde todos os três nucleotídeos foram alterados (fig. 4).

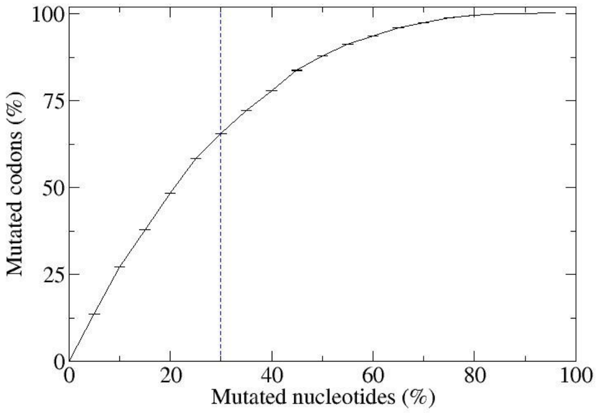


Figura 4. Porcentagem media dos códons mutados como função da porcentagem dos nucleotídeos mutados para o experimento de mutação randômica.

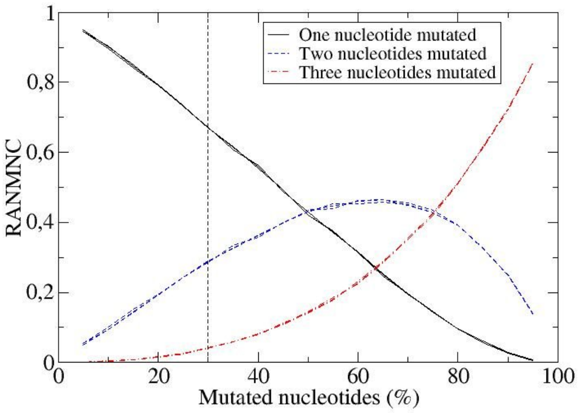


Figura 5. Porcentagem media dos códons com um, dois e três nucleotídeos mutados como função da porcentagem de nucleotídeos mutados em um ciclo de mutação randômica. O eixo vertical RANMNC significa: número médio relativo de nucleotídeos mutados por códon.

Discussão

A AMI tem capacidade de analisar seqüências de DNA através da periodicidade três correspondente a estrutura de códons. Essa característica indica que a entropia criada por essa estrutura é responsável pela medida codificante (VAIDYANATHAN et al, 2002). A AMI mede, assim, o potencial codificante de uma seqüência. Esse potencial varia bastante entre as seqüências codificantes, chegando até mesmo a confundir algumas CDS com seqüências não-codificantes, sendo que os casos que mais ocorrem isso são em organismos mais complexos como Homo Sapiens. Isso sugere que esse potencial foi alterado durante a evolução através, principalmente, das mutações. É muito provável que tenha ocorrido uma compensação da perda de entropia na organização dos códons pela introdução de novas estruturas no gene desses organismos como o *splicing* e mecanismos de regulação mais complexos.

Como esperado, o nível de mutações in silico introduzidas pelo procedimento descrito foi o mesmo nas três repetições do mesmo experimento. A variação de um para o outro foi no máximo de 3%. Portanto, a medida de codificação que é perdida em alguns casos com apenas 15% de mutações não pode ser o resultado de um caso trivial de uma distribuição de mutações voltadas para posições conservadas dos códons.

O tamanho das seqüências utilizadas foram superiores a 350 bp já que seqüências menores têm um decréscimo na quantidade de informação devido a pouca representatividade estatística Na perspectiva de medida de entropia da seqüência codificante usando a AMI, toda a medida codificante de uma seqüência se perde alterando apenas 30% da seqüência. Entretanto, 15% de mutações já são suficientes para destruir a característica codificante de algumas seqüências indicando provavelmente que essas seqüências já perderam uma parte do poder codificante e não apresentaram seu máximo potencial codificante. Dessa forma as seqüências podem ser classificadas pelo seu potencial codificante extraindo informações contidas em sua correlação espacial entre códons. O próximo passo será investigar se esses agrupamentos podem ser justificados por uma função biológica. Nesta hipótese, pode-se inferir que seqüências distantes do valor biológico esperado podem ter acumulado algumas mutações. Um caso extremo desse tipo é conhecido como pseudogenes. Se existir uma correlação entre o valor da medida codificante e a presença de alguma função biológica ela terá uma grande importância para investigações de evolução gênica.

Agradecimentos

Esta pesquisa foi suportada por agencias brasileiras CNPq e FAPESB que financiaram os pesquisadores N.C. e R.V., respectivamente.

Referencias

ABU-HANNA, A. (1999) Review of "Machine Learning," Artificial Intelligence in Medicine, Elsevier, 16, 201-204

ALTSCHUL, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. J. Mol. Biol. 215, 403-410.

BURGE, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268, 78-94.

CARELS, N., Vidal, R., Mansilla, R. and Frias, D. (2004) The mutual information theory for the certification of rice coding sequences. FEBS Lett. 568,155-8.

GROSSE, I., Buldyrev, S.V. and Stanley, H.E. (2000) Average Mutual Information of Coding and non-coding DNA. Pacific symposium on Biocomputing 5, 611-620.

KULP, D., Haussler, D., Reese, M.G. and Eechman, F.H. (1996) A generalized hidden Markov model for the recognition of human genes in DNA. Proc. Int. Conf. Intelligent. Syst. Mol. Biol. 4, 134-142.

LI, W.; Marr, T.G.; Kaneko, K. (1994) Understanding long-range correlations in DNA sequences. Physica D. V. 75, p. 392–416.

MATHÉ, C., Sagot, M.F., Schiex, Th. and Rouzé, P. (2002) Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res. 30, 4103-4117.

SALZBERG, S.L., Pertea, M., Delcher, A.L., Gardner, M.J. and Tettelin, H. (1999) Interpolated Markov models for eukaryotic gene finding. Genomics 59, 24–31.

SHANNON, C.E. (1948) A mathematical theory of communication. Bell Systems Technical Journal V. 27, p. 379-423.

TIWARY, S. Ramchandran, S., Bhattacharya, A., Bhattacharya, S. and Ramaswamy, R. (1997) Prediction of probably genes by Fourrier analysis of genomic sequences. CABIOS 13, 263-270.

VAIDYANATHAN, P. P. and Yoon, B., (2002) Digital Filters for Gene Prediction Applications. Proc. 36th Asilomar Conference on Signals, Systems, and Computers, Monterey, CA.

XU, Y., Mural, R.J., Einstein, J.R., Shah, M.B. and Uberbacher, E.C. (1996) Grail: a multiagent neural network system for gene identification. Proc. IEEE 84, 1544-1551.

4. CAPITULO 2

FRAMER: A method of reading frame detection based on joint frequency of nucleotide pairs

Autores: **Ramon Vidal**, Nicolas Carels, Diego Frias.

Instituição: Laboratorio de Bioinformatica, Universidade Estadual de Santa Cruz, Rodovia Ilheus/Itabuna km. 16, Ilhéus, Bahia, Brazil.

Submetido: DNA Research.

FRAMER: A method of reading frame detection based on joint frequency of nucleotide pairs

Ramon Vidal, Nicolas Carels* and Diego Frías

Laboratório de bioinformática, Universidade Estadual de Santa Cruz, Rod. Ilhéus/Itabuna km. 16, 45662-000 Ilhéus, Bahia, Brazil.

ABSTRACT

Motivation: ESTs are often contaminated by non-coding stretches. Therefore, coding frame cannot be obtained with absolute certainty when ESTs do not have homology with known coding sequences. Below 400 bp, codon stop statistics is often ambiguous as well as codon usage when searching for most probable reading frame. Therefore, an efficient tool of reading frame detection independent from biological species, robust to sequence size and non-coding stretches is needed for proper EST characterization.

Results: In simple eukaryotes, the method was able to find the correct reading frame of coding sequences >50 bp and >250 bp with efficiencies >90% and >95%, respectively. In insects and vertebrates, the efficiency was reduced by ~5%. Our coding sequences were experimentally proven and covered the full range of GC level in third position of codons.

Contact: admin@labbi.uesc.br

Supplemental information: Data comprising the benchmarks used in the paper may be downloaded from http://labbi.uesc.br/FRAMER/

INTRODUCTION

Partial sequencing of randomly-selected cDNA clones, i.e. expressed sequence tags (ESTs), remains essential in the process of gene identification and transcriptome characterization. In eukaryotes, gene detection by computing methods generally needs to be assisted by EST homology search (Reese *et al*., 2000; International Rice Sequencing Consortium, 2005) in an effort to detect false positive genes (Cruveiller *et al*., 2004; Carels *et al*., 2004). As a consequence, EST projects remain an essential step for genome annotation, but also often for analyses of gene expression relying on micro-arrays, SAGE or MPSS (Reinartz *et al*., 2001).

The coding information of an EST can be corrupted with chimerical insertions, unspliced messenger, UTR, random sequences and/or heterologous sources. Most methods proposed for solving this problem are codon usage (Sharp *et al*., 1988) dependent (Fickett, 1982; Staden and McLachlan, 1982; White *et al*., 1993). This has two consequences: first, their precision depends on the size and quality of the training set; second, the sample of sequences detected can be biased because of overfitting (Abu-Hanna, 1999). These methods together with complex Hidden Markov Models (Krogh *et al*., 1994; Baldi and Brunak, 2001) or neural networks (Lapedes *et al*., 1990; Uberbacher and Mural, 1991; Farber *et al*., 1992; Xu *et al*., 1994; Snyder and Stormo, 1995) are not longer usable when new biological species are analyzed because the model on which they are based is not defined yet. This is an obvious limitation for biodiversity investigations. In addition, in interaction libraries such as those prepared with host tissues infected by a parasite or a symbiont, the probability model of the host can be partly compatible with that of the parasite and *vice versa* leading to produce significant rate of false positives and negatives.

Due to their organization in codons that introduce a periodicity of order 3, coding sequences (CDS) are characterized by a lower entropy level compared to non-coding DNA (see Wang *et al*., 2004 for a review). Mathematical models based on Discrete Fourier Transform (DFT) (Tiwary *et al*., 1997; Li, 1997; Fukushima *et al*., 2002; Kotlar and Lavner, 2003), Average Mutual Information (Grosse *et al*., 2000), nucleotide statistics (Zcurve) (Guo *et al*., 2003) and ancestral codon (Nikolau and Almirantis, 2004) were elaborated to detect this periodicity. The features that detect the periodicity introduced by codons in the DNA are universal to the biosphere so as it is for the genetic code, therefore, largely independent of the codon usage. However, these methodologies are not able to identify the correct open reading frame contrary to those that rely on codon usage (Neural Networks and Markov models). The only exception is the method proposed by Anastasiou (2000), which is based on the ratio of linear combinations of nucleotide autocorrelations. However, the discriminator proposed by Anastasiou (2000) needs to be "tuned" through the optimization of weighting factors using a training set of CDS. Because of this, the method is, finally, also strongly dependent on codon usage. Zcurve (Guo *et al.* 2003) only succeeds in classifying a CDS if it is already in frame. Otherwise, if the sequence is shifted by one or two bases it will produce a false negative. In the method based on the ancestral codon RNY (Nikolau and Almirantis, 2004) the maximum of the function CSF that measures the asymmetry of codon frequency is used to detect whether a sequence is coding or not. This function does not detect the coding frame. Therefore, these methods neither apply to the case of coding frame detection irrespective of the base composition of the sequence analyzed.

It was also observed that purines (adenine or guanine) occur predominantly at the first position of codons (Shepherd, 1981; Fickett, 1982; Staden, 1984; Yan *et al*., 1998; Majumdar *et al*., 1999; Eskesen *et al*., 2004). However, this feature was not systematically used for reading frame detection in complicated sequence contexts.

It is the purpose of this article to show how to perform purines and other nucleotides counting for reading frame detection in ESTs.

* To whom correspondence should be addressed.

Considering (i) a DNA sequence divided in *Nc* triplets, (ii) the *joint frequency* (F=f_if_j) as the product of the relative frequencies (f_i or f_j =(N_i or N_j)/$3N_c$) of the two nucleotides (N_i or N_j = A,C,G or T) of a pair, and (iii) the position p (1,2 or 3) of the triplets for which the value of F is maximum (F=Φ), the method (FRAMER) sets that the position p of the triplets where Φ is found corresponds to the coding frame, i.e. Φ is in position +1 of codons.

Given the actual state of the technology of DNA sequencing, the robustness of Φ can be characterized by its ability to detect the correct reading frame for sequences between 50 and 600 bp eventually contaminated by non-coding stretches and stop codons. In addition, the predictive efficiency of Φ must be acceptable independently of the GC level and the codon usage. We estimated the robustness of Φ by the ratio of its average efficiency to the difference of its maximum and minimum efficiency for CDS between 50 and 600 bp. With such concern, we noticed that FRAMER is (i) less efficient in insects and vertebrates by 5% than in any other organisms (prokaryote or eukaryote), (ii) not significantly affected by small non-coding stretches (below 100% of the coding stretch) and (iii) not affected by in frame stop codons that often occur due to sequencing errors or any type of non-coding sequences. The sensitivity of Φ was compared to that of reading frame detection by local alignments (BLASTX) scoring the ratio of false positives to true positives.

METHODS

2.1 Coding sequence samples

We tested coding sequences (CDS) from *Arabidopsis thaliana* (GC3 = 25 to 65%; Carels *et al*., 1998), *Plasmodium falciparum* (GC3 = 0 to 30%; Musto *et al*., 1995), and *Chlamydomonas reinhardtii* (GC3 = 60 to 100%; Naya *et al*., 2001; Scala *et al*., 2002) because they cover the complete range of GC (guanine + cytosine) level in third position of codons (GC3) and extreme codon usage found in eukaryote CDS. In addition, we tested *Oryza sativa* (GC3 = 25 to 100%; Carels *et al*., 2004), because the CDS of this species contain the complete GC range observed in *A. thaliana* and *C. reinhardtii*, taken together. Finally, we also tested CDS from *H. sapiens* and *D. melanogaster* considered, here, representative for vertebrates and insect lineages. The CDS sample size of all these species (6 datasets) is statistically representative and the species themselves are model systems in genetics.

Complete nuclear CDS of the above species were retrieved from GenBank (release 137, August 15th, 2003) and filtered according to Carels *et al*. (2004) to eliminate redundancy and genes from *in silico* predictions.

2.2 The sensitivity of Φ to the coding sequence size

We prepared 12 datasets according to the CDS size for each of the 6 species. Each dataset differed by a sequence size of 50 bp within the range of 50 bp to 600 bp. The dataset of a given sequence size was obtained by cutting off the supernumerary nucleotides from the 3' side of the original CDS. The sensitivity was estimated by the relationship between the rate of false positive generated by Φ and the CDS size.

2.3 The efficiency of coding frame detection of Φ

For testing the efficiency and the sensitivity of coding frame detection in practical applications involving ESTs, we built a dataset of nucleotide sub-sequences of the japonica subset of GenBank (release 137, August 15th, 2003) EST database (123,127) homologous (BLASTX) to the complete set of rice CDS from TIGR (version 1.0) that we previously cleaned for false positives (Carels *et al*., 2004) and translated into amino acid sequences (24,914). The 985 EST homologous sub-sequences > 399 bp with expected value E < 0.0001 obtained by this process were necessary coding and in frame. These sequences were used as a control to measure the coding frame detection efficiency of Φ.

2.4 The coding frame and the statistics of stop codons

It is common practice to search the frame with the lowest stop codon frequency and to set it as the reading frame. Therefore, we counted the number of stop codons appearing in their 3 frames (+1, +2 and +3) for each homologous sub-sequence detected with BLASTX (paragraph 2.3) and estimated the power of this methodology for coding frame detection. The rate (%) of false positives in coding frame detection by stop codon statistics was compared to that of BLASTX and Φ in the practical case of ESTs.

2.5 The robustness of Φ to non-coding DNA

ESTs are generally made up of a coding stretch flanked by non-coding sequences typically in a range below 30% of the coding stretch. Therefore, we tested the robustness of the efficiency of coding frame detection with Φ by adding flanking non-coding (random) stretches to CDS of all datasets of this study. The sequences size of these non-coding sequences varied between 10% and 400% of that of the original CDS.

2.6 The reading frame prediction with Φ

Let $\Gamma = \{(i,j),\ i,j \in \{A,C,G,T\}\}$ denote the set of 10 different duplets of nucleotides, without concern with the order of nucleotides in the duplet and σ be the set of all different combinations of n = 1,…,10 duplets in Γ so that each element, σ_k, k=1,…,1022, represents a n-tuplet of pairs of nucleotides.

The joint frequency of a DNA sequence at codon position p=1,…,3 for the *k-th* combination of duplets is defined as

$$F_{\sigma_k}^{p} = \sum_{(i,j)\in\sigma_k} f_i^{\,p} f_j^{\,p}$$

where $f_i^{\,p}$ is the frequency of the nucleotide i at codon position p relative to the number of codons (triplets) in the DNA sequence.

For a small subset of σ it was observed that the maximum value of F_{σ}^{p}, lets denote it by Φ, is found in position +1 of codons, i.e. Φ is an indicator of the coding frame.

In order to compare the different combinations, we introduced: (i) the average efficiency, *m,* (ii) the difference (*d*) between the maximum and minimum efficiency (e) observed for all range of sequence size, $d = max(e) – min(e)$, and (iii) a robustness index $r = m / (1 + d)$.

Φ is designed to detect coding frame of sequences that are known to be coding and are on the leading strand (5' to 3').

With the aim to obtain a reference for the robustness index, we investigated the efficiency of BLASTX (version 2.2.8, January 5th, 2004) in predicting the reading frame of the sequences. BLASTX is able to predict the reading frame by finding the best alignment of the CDS fragments with homologous sequences in the non-redundant protein database (*nr*) of GenBank. In this study we used the GenBank release 141, April 14th, 2004, and the BLOSUM62 matrix for BLASTX with maximum E-value = 0.0001.

All the plots were obtained with Grace-5.1.18 (http://plasma-gate.weizmann.ac.il/Grace/).

RESULTS

The reading frame detection considering **Φ** of A and G only, performed very well in eukarotes other than insects and veretebrates. In fact, the efficiency (rate of correct predictions) of the method (FRAMER) was not signifantly dependent either on the GC level of the sequences or on the codon usage, as shown in Table 1.

We found similar performances of **Φ** for species covering all the GC3 range of eukaryotes with an average efficiency of 95.3% (standard deviation: 3.6%). The best prediction level (98.6%) was obtained with *Plasmodium falciparum* which is extremely GC-poor, but FRAMER was also performing very well (95.1%) with *Chlamydomonas reinhardtii* (GC-rich) and rice (*Oriza sativa)* (94.2%), in both GC-poor and GC-rich genes (Carels and Bernardi, 2000; Carels *et al*., 2004; Carels, 2005).

Looking for the optimal sets of nucleotide pairs to be used in the calculation of **Φ**, the complete set of 1,022 combinations of 10 permutable nucleotide pairs was evaluated with data sets of different sequence size, ranging from 50 bp to 600 bp. We found that the 10 best combinations, considering the efficiency, contain the nucleotide pair (A,G) (Table 2), but the robustness can be improved when additional nucleotides are taken into account (Table 3). We also noted that the pairs of nucleotides (C,C), (T,T) and (T,C) contribute only a little and are dominant in the combinations with the poorest performances (data not shown).

Examining the values of *d*, *m* and *r* for each combination of nucleotide pairs, we noticed that for the species with highest compositional homogeneity (*P. falciparum, A. thaliana* and *C. reinhardtii*), the best values of *d*, *m* and *r* are found for one combination of nucleotide pairs. Consequently, the combination of nucleotide pairs with the highest performance of reading frame detection (highest *m* and *r* values) is also the most robust to CDS size variation (lowest *d* value). By contrast, in species with higher compositional heterogeneity (*O. sativa*, *D. melanogaster, H. sapiens*), we found that robustness and performance are no longer linked and a compromise between these two factors must be searched when fine tuning is desired. The best compromise is obtained for the highest *r* value, especially below 250 bp. Above 250 bp, robustness does not matter very much and the variable can be chosen according to the best value of *m* (Figure 1).

The best average performance was reached with the combination number 5, yielding an average efficiency over species of 92.1%, with a variation with sequence size of only 8% (Table 2). Compared with combination 10, corresponding to the nucleotide pair A,G, we noted an improvement of almost 2% in efficiency and of 100% in robustness. It is interesting to stress here that except for *H. sapiens* and *D. melanogaster*, the prediction efficiency of combinations 5 and 10 were similar.

Table 1. Prediction efficiency (%) of reading frame detection with joint frequency of A and G for species covering all the GC3 range of eukaryotes. The coding sequences originally in frame (FB1), were shifted by removing one (FB2) or two (FB3) nucleotides from the 5' side. The predicted coding frame (+1) is in bold and appears at different positions *p* (‡) of the *putative codon* depending on the actual position of the first base of the codon of the sequence. The non-bold percentages are for the rate of false positives. Rice is made up of two gene populations (*): a GC3-poor (45%) and a GC3-rich (95%). N is for sample size.

			FB 1			FB 2			FB 3		
Species	**N**	**GC3 (%)**	***1*‡**	*2*	*3*	*1*	*2*	***3***	*1*	***2***	*3*
P. falciparum	209	*15*	**99.53**	0.00	0.47	0.47	0.00	**99.53**	0.00	**99.53**	0.47
A. thaliana	1,323	*45*	**96.67**	0.61	2.72	3.34	0.90	**95.76**	0.00	**95.76**	3.34
C. reinhardtii	114	90	**98.24**	0.00	1.76	1.76	0.00	**98.24**	0.00	**97.37**	2.63
O. sativa	442	*45+95**	**95.02**	0.46	4.52	3.65	1.37	**94.98**	1.16	**94.21**	4.63

Table 2. Combinations of the 10 nucleotide pairs with highest prediction efficiencies for coding frame detection. When '1' is associated to a nucleotide pair (i,j), it means that it is used for reading frame prediction, otherwise it is not. # is for the combination number.

# \ i,j	A,G	A,C	T,G	G,G	A,A	C,G	A,T	C,C	T,T	T,C
1	1	1	1	1	1	0	0	1	1	0
2	1	1	1	1	1	1	1	0	0	0
3	1	1	1	1	0	0	1	1	0	0
4	1	1	1	1	1	0	0	1	0	0
5	1	1	1	0	0	1	1	0	0	0
6	1	1	1	0	1	1	0	0	0	0
7	1	1	0	1	0	0	0	0	1	0
8	1	1	1	1	0	0	0	0	0	0
9	1	1	0	1	0	0	0	0	0	0
10	1	0	0	0	0	0	0	0	0	0

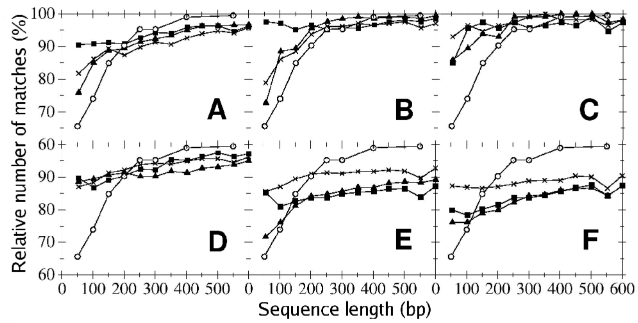


Fig. 1. Variation of the efficiency of FRAMER with the sequence length for three of the best combinations of nucleotide pairs listed in Table 2, That is, (■) 2, (**X**) 5 and (▲) 10, for six species (A) *A. thaliana*, (B) *P. falciparum*, (C) *C. reinhardtii* (D) *O. sativa*, (E) *D. melanogaster*, and (F) *H. sapiens*. The homology search between rice coding sequences against *nr* (GenBank) with BLASTX (**O**) was used as control of prediction efficiency.

Table 3. Performance of the best 10 combinations of nucleotide pairs listed in Table 2 for various species. Concerning the performance of reading frame prediction for coding sequences size varying between 50 and 600 bp (i) *d* is the difference between the minimum and the maximum of prediction efficiencies across the sequence size range, (ii) *m* is the average efficiency and (iii) *r* is the robustness index given by *m /(1 + d)*. The maximum *r* value represents the best compromise for average efficiency and robustness to sequence size. The combinations that are (i) the most robust are in bold, (ii) the best for average efficiency are in bold and underlined and (iii) the best for *r* are in bold and italic. The values involved in best scores are printed on gray background. # is for the combination number.

	P. falciparum			*A. thaliana*			*C. reinhardtii*			*O. sativa*			*D. melanogaster*			*H. sapiens*			*Average over Species*		
#	d	m	r	d	m	r	d	m	r	d	m	r	d	m	r	d	m	r	d	m	r
1	21.5	95.4	4.2	9.3	91.2	8.9	7	96.3	12	5.9	93.2	13.5	**4.1**	**87.7**	***17.2***	4.2	87.3	16.9	8.7	91.9	12.1
2	**3.4**	**96.8**	***22.3***	**5.8**	**93.7**	***13.7***	14	95.7	6.4	10.6	**93.3**	8	6.3	84.7	11.5	9.3	83.7	8.1	8.2	91.3	11.7
3	18.7	93.4	4.8	14.5	89	5.8	7	97.2	12.1	9.3	93.2	9.1	7.2	**90.5**	11	**3.9**	**88.4**	***18.2***	10.1	92.0	10.2
4	15.8	95.2	5.7	19	90	4.5	6.2	97.9	13.7	9.5	92.9	8.9	9.9	86.9	8	7.7	88.3	10.2	11.4	91.9	8.5
5	19.6	90.8	4.4	8	94	10.4	**6.2**	**98**	***13.7***	**3.8**	**90.3**	***18.6***	5.1	90.4	14.9	5.2	89.2	14.5	**8.0**	**92.1**	***12.8***
6	14.4	91.5	6	10.2	92.1	8.2	8.8	97.9	10	5.2	91.6	14.8	9.2	86.9	8.6	6.1	88.7	12.5	9.0	91.5	10.0
7	45.9	86.7	1.8	20.3	91.5	4.3	11.4	96.4	7.8	17.2	92.2	5.1	19.8	85.6	4.1	16.5	84.1	4.8	21.9	89.4	4.7
8	36.4	92.3	2.5	26.8	88.9	3.2	21.9	93.2	4.1	18.6	92.7	4.7	25.8	83.4	3.1	19.7	**93.1**	4.5	24.9	90.6	3.7
9	26.8	93.6	3.4	19	91.2	4.5	9.7	97.2	9.1	20.4	92	4.3	20.4	86	4	18	85.3	4.5	19.1	90.9	5.0
10	26.8	94.7	3.4	20.7	91.4	4.2	14	96.3	6.4	6.3	91.6	12.5	17.5	84.4	4.6	11.1	82.8	6.8	16.1	90.2	6.3

We found that FRAMER gives higher rate of true positives in *A. thaliana*, *O. sativa*, *C. reinhardtii* and *P. falciparum* than in *D. melanogaster* and *H. sapiens* (Fig. 1) since the average efficiency was, in these latter species, 5% to 10% below the former, but still higher than 90% for combination 5. Figure 1 also shows that in the hypothesis that homologous sequences are available, BLASTX performs better than FRAMER for sequence size higher than 200 bp. However, for shorter sequences the BLASTX algorithm does not always succeed to find a homologous sequence, while FRAMER still has detection efficiency 10% higher than BLASTX at 100 bp. At 50 bp the difference of efficiency between both methods reaches 20-30% (Fig. 1). The loss of efficiency as a function of sequence size reduction is almost linear for FRAMER, but is exponential for

BLASTX. For some nucleotide combinations, the efficiency was found little affected by CDS size (2 in *A. thaliana* and *P. falciparum*, 5 in *C. reinhardtii* and *O. sativa*, 1 in *D. melanogaster* and 3 in *H. sapiens*). The species for which the efficiency was found most size dependent was *P. falciparum* with up to 25% for the combination 10 (Fig. 1). However, in this specific case combination 2 gave excellent results with variation of only 3.4%, (Table 3). We also found that BLASTX is the best method for reading frame prediction when CDS are larger than 200 bp except for species with extreme compositional bias such as *P. falciparum* and *C. reinhardtii* where FRAMER always performed better or equally well (combinations 2 and 3, respectively). For the other species, the efficiency of FRAMER was about 5% lower than BLASTX when CDS size was >300 bp except for *D. melanogaster* and *H. sapiens* where the performance loss reached 15% (Fig. 1).

In the dataset of EST homologous sub-sequences (985 sequences or 193,888 codons), we found a total of 88 stop codons in frame +1, 6,410 in frame +2 and 8,520 in frame +3. In most cases only one in-frame stop codon was found generally in the middle of the ORF. FRAMER identified the EST reading frame with an efficiency of 98.27%, detecting the correct frame even with the presence of non-coding stretches or in-frame stop codons.

Attempting to detect the reading frame of such ESTs by looking for the minimum number of in-frame stop codons, we noted an error rate of 16.24%. The weakness of this approach is mainly due to the occurrence of the same stop codon frequency in two frames, introducing ambiguity as to which is the correct one.

Concerning, the ability of FRAMER to correctly predict the coding frame when it is flanked by non-coding sequences, we found that the rate of detection efficiency decreased by <1.5% when the flanking random sequences were in the range of 10% to 100% of the coding stretches size and by 3% to 4% when random sequences were in the range of 200% to 400% of the coding stretch size, respectively.

DISCUSSION

As seen above, the level of detection of coding frame by **Φ** with combinations of nucleotide pairs including A and G is the highest at the first position of the codon. This allows the coding frame detection with a very high tolerance level to codon usage variations independently of the codon position of the 1st nucleotide of the sequence, which makes it useful for reading frame detection in ESTs of organisms for which GC level and/or codon usage were not identified yet. In addition it shows that the predominance of purines in first position of codons is a powerful feature for coding frame detection. This feature also indicates the coding status of a sequence and the coding strand (data not shown), but it is not the purpose of this article to discuss **Φ** in the contexts of (i) coding/non-coding diagnostic, (ii) coding strand diagnostic, or (iii) gene finding process. Such considerations will be presented elsewhere.

The fact that the efficiency of reading frame prediction with **Φ** is robust to the codon usage when calculated with the nucleotide pair A and G suggests a compensation between these two nucleotides in eukaryotes that are biased toward AT or GC such as *P. falciparum* and *C. reinhardtii*, respectively. A and G are both purines and the preference of a purine in 1st codon position is, therefore, an universal feature of coding sequence (CDS) since it is conserved throughout eukaryotes and prokaryotes (data not shown), irrespective of GC range. The other nucleotides C and T have only an additive effect on coding frame detection (Table 2).

Since GC level and codon usage are not responsible for the lower performance (5-10%) of the method (FRAMER) in *H. sapiens* and *D. melanogaster*, we concluded that it is due to the higher entropy of CDS of these species. This conclusion is sustained by the fact that combination 5 that implements a mixture of purines and pyrimidines gives higher rate of prediction in these species. The constraint of purine use at the first position of codons is slightly lower in *D. melanogaster* and *H. sapiens* than in the other species of this study. The relationship between this observation and organism complexity, if any, will be investigated elsewhere. However, it appears clearly form Table 3 that the prediction efficiency of the pair A,G (combination 10) decreases regularly with the increase in species complexity.

The low sensitivity level of **Φ** to the actual position of the first nucleotide of the codon in the analyzed sequence is due to the fact that it is contextual. The linear relationship of **Φ** with the CDS size shows that the lower efficiency at small CDS size is due to increasing lack of statistical significance. Because of this, the discrimination efficiency of exons and introns by AMI (Grosse *et al*., 1999; 2000) vanishes below 400 bp. This is not interesting for EST investigation given the relatively small size of their CDS stretch with satisfactory quality levels (Phred quality > 10). Regarding reading frame detection, **Φ** is poorly dependent on CDS size because it does not rely on the choice between coding and non-coding sequences as it is the case with AMI. The comparison of $F^{(p)}$ among codon positions is internal to the same context, thus, it is not the signal intensity associated with reading frame detection that is important, it is rather the rate of the noise increase with the CDS size reduction. This rate is linear in the case of **Φ** and much lower that the loss of homology detection of BLASTX that is exponential for sequences < 200 bp.

The fact that the robustness of **Φ** is lower in compositionally heterogeneous species (*O. sativa, D. melanogaster, H. sapiens*) can be understood intuitively from the fact that the increase of entropy introduces noise in the periodicity because of the higher heterogeneity of nucleotide combinations. In that sense, rice is interesting because it is made up of only two classes of genes each one homogeneous in comparison to the higher continuity level of heterogeneity in *D. melanogaster* and *H. sapiens* (Paces *et al*., 2004; Carels, 2005; Constantini *et al.*, 2006). It is probably the reason why FRAMER performs better in rice than in *D. melanogaster* and *H. sapiens*.

The idea of searching the frame with the fewest or no stop codons to find out the reading frame is not realistic because of the significant occurrence of in-frame stop codons in this material. BLASTX and FRAMER can cope with this situation. It is worthwhile to stress here that looking for robustness we analyzed the stop codon frequencies in the three frames of sequences as large as 600 bp. Actually, the informative part of ESTs is often shorter, typically in the range of 350 to 400 bp. In addition, in-frame stop codons from 5' and 3' UTR, non-spliced introns and random sequences may introduce additional noise.

These contexts are not included in our experiment because local alignment eliminates them. The ambiguity of reading frame determination based on stop codon statistics increases exponentially with CDS size reduction. Methods based on stop codon frequency perform poorly within this size range (data not shown).

The prediction efficiency of coding frame CDS is little affected by the presence of non-coding stretches until 100% of their size. This is expected since the contribution of random sequences to **Φ** is null due to their higher entropy. FRAMER is, therefore, ideal for removing remnant non-coding sequences after the extraction of the sequence piece of acceptable Phred quality. The detection of in frame stop codons that surround the coding part of the EST allows further cleaning process.

Finally, **Φ** does not require any training procedure (such as it is for Hidden Markov Model and neural network) or preexisting information (such as it is for BLASTX) to be efficient. Its rather good performances even for short CDS makes it useful for reading frame detection in EST processing. These features make it an attractive tool for EST projects with low level of preexisting information. BLASTX is very efficient for correct reading frame detection when homologous sequences are available. The error rate of BLASTX for reading frame detection is <1% irrespective of the sequence length. Small CDS sizes are common when extraction of high quality mRNA is difficult to obtain. This situation often occurs when new species or tissues are taken into consideration. In such cases FRAMER may appear useful, since the efficiency of homology detection by BLASTX drops exponentially for CDS smaller than 200 pb.

ACKNOWLEDGEMENTS

We thank Carter Miller and Martin Brendel for the manuscript revision. This research was supported by the Brazilian agencies CNPq and FAPESB providing researcher fellowships to N. Carels and student fellowships to R. Vidal.

REFERENCES

Abu-Hanna,A. (1999) Review of "Machine Learning," Artificial Intelligence in Medicine, Elsevier. 16: 201-204.

Anastasiou,D. (2000) Frequency-domain analysis of biomolecular sequences. *Bioinformatics* 16: 1073-1081.

Baldi,P. and Brunak,S. (2001). Bioinformatics: The machine learning approach 2nd ed., chapter 7. MIT Press, Cambridge, MA.

Carels,N. and Bernardi,G. (2000). Two classes of genes in plants. *Genetics* 154: 1819-1825.

Carels,N., Hatey,P., Jabbari,K., and Bernardi,G. (1998) Compositional properties of homologous coding sequences from plants *J. Mol. Evol.* 46: 45-53.

Carels,N., Vidal,R., Mansilla,R., and Frías,D. (2004) The mutual information theory for the certification of rice coding sequences. *FEBS Lett.* 568: 155-158.

Carels,N. (2005) The genome organization of angiosperms. In: Recent Research Developments in Plant Science. Ed: Pandalai, S.G., Research Signpost, Trivandrum, Kerala, India, pp. 129-194.

Constantini,M., Clay,O., Auletta,F. and Bernardi,G. (2006) An isochore map of human chromosomes. *Genome Res.* 16: 536-541.

Cruveiller,S., Jabbari,K., Clay,O., and Bernardi,G. (2004) Incorrectly predicted genes in rice? *Gene* 333: 187-188.

Eskesen,F.N., Kihghorn,B., and Ruvinsky,A. (2004) Periodicity of DNA in exons. *BMC Mol. Biol*. 5: 12.

Farber,R.B., Lapedes,A.S., and Sirotkin,K.M. (1992) Determination of eukaryotic protein coding regions using neural networks and information theory. *J. Mol. Biol.* 226: 471-479.

Fickett,J.W. (1982) Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res*. 10: 5303-5318.

Fukushima,A., Ikemura,T., Kinouchi,M., Oshima,T., Kudo,Y., Mori,H., and Kanaya,S. (2002) Periodicity in prokaryotic and eukaryotic genomes identified by power spectrum analysis. *Gene* 300: 203-211.

Gouy,M., Gautier,C., Attimonelli,N., Lanave,C., and Di Paola,G. (1985) *CABIOS* 1: 167-172.

Grosse,I., Herzel,H., Buldyrev,S.V., and Stanley,H.E. (1999) Species independence of mutual information in coding and non-coding DNA. *Phys. Rev.* E 61: 5624-5629.

Grosse,I., Buldyrev,S.V., and Stanley,H.E. (2000) Average Mutual Information of coding and non-coding DNA. *Pacific symposium on biocomputing* 5: 611-620.

Guo,F.-B., Ou,H.-Y. and Zhang,C.-T. (2003) Zcurve: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. *Nucleic Acids Res.* 31: 1780-1789.

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature* 36: 793-800.

Kotlar,D. and Lavner,Y. (2003) Gene prediction by spectral rotation measure: a new method for identifying protein-coding regions. *Genome Res.* 13: 1930-1937.

Krogh,A., Mian,I.S., and Haussler,D. (1994) A hidden Markov model that finds genes in E. coli DNA. *Nucleic Acids Res*. 22: 4768-4778.

Lapedes,A.S., Barnes,C., Burks,C., Farber,R.B., and Sirotkin,K.M. (1990) Application of neural networks and other machine learning algorithms to DNA sequence analysis. In Computers and DNA. Eds. G. Bell. and T. Marr., Addison-Wesley, Redwood City, CA. pp. 157-182.

Li,W. (1997) The study of correlation structures of DNA sequences: a critical review. *Computers Chem.* 21: 257-271.

Majumdar,S., Gupta,S.K., Sundararajan,V.S., and Ghosh,T.C. (1999) Compositional correlation studies among the three different codon positions in 12 bacterial genomes. *Bioch. and Bioph. Res. Com.* 266: 66-71.

Musto,H., Rodriguez-Maseda,H. and Bernardi,G. (1995) Compositional properties of nuclear genes from Plasmodium falciparum. *Gene* 152: 127-132.

Naya,H., Romero,H., Carels,N., Zavala,A., and Musto,H. (2001) Translational selection shapes codon usage in the GC-rich genome of Chlamydomonas reinhardtii. *FEBS Lett.* 501: 127-130.

Nikolaou,C. and Almirantis,Y. (2004) Measuring the coding potential of genomic sequences through a combination of triplet ocurrence patterns and RNY preference. *J. Mol. Evol*. 59: 309-316.

Paces,J., Zika,R., Paces,V., Pavlicek,A., Clay,O. and Bernardi,G. (2004) Representing GC variation along eukaryotic chromosomes. *Gene* 333: 135-141.

Reese,M.G., Kulp,D., Tammana,H. and Haussler,D. (2000) Genie-gene finding in Drosophila melanogaster. *Genome Res*. 10: 529-538.

Reinartz,J., Bruyns,E., Lin,J.-Z., Burcham,T., Brenner,S., Bowen,B., Kramer,M., and Woychik,R. (2002) Massively parallel signature sequencing MPSS. as a tool for in-depth quantitative gene expression profiling in all organisms. *Brief. Funct. Gen. Prot.* 1: 95-104.

Scala,S., Carels,N., Falciatore,A., Chiusano,M.L. and Bowler,C. (2002) Genome properties of the diatom *Phaeodactylum tricornutum*. *Plant Physiol.* 129: 993-1002.

Sharp,P.M., Cowe,E., Higgins,D.G., Shields,D.C., Wolfe,K.H. and Wright,F. (1988) Codon usage patterns in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*: a review of the considerable within-species diversity. *Nucleic Acids Res.* 16: 8207–8211.

Snyder,E.E. and Stormo,G.D. (1995) Identification of protein coding regions in genomic DNA. *J. Mol. Biol*. 258: 1-18.

Staden,R. (1984) Measurements of the effects that coding for a protein has on a DNA sequence and their use for finding genes. *Nucleic Acids Res*. 12: 551-567.

Shepherd,J.C.W. (1981) Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. *Proc. Natl. Acad. Sci.* USA 78, 1596-1600.

Staden,R. and McLachlan,A.D. (1982) Codon preference and its use in identifying protein coding regions in long DNA sequences. *Nucleic Acids Res.* 10: 141-156.

Tiwary,S., Ramchandran,S., Bhattacharya,A., Bhattacharya,S. and Ramaswamy,R. (1997) Prediction of probably genes by Fourier analysis of genomic sequences. *CABIOS* 13: 263–270.

Trifonov,E.N. (1998) 3-, 10.5-, 200- and 400-base periodicities in genome sequences. *Phys. A*. 249: 511-516.

Uberbacher,E.C. and Mural,R.J. (1991) Locating protein-coding regions in human DNA sequences by a multiple senso-neural network approach. *Proc. Natl. Acad. Sci.* 88: 11261-11265.

Wang,Z., Chen,Y. and Li,Y. (2004) A brief review of computational gene prediction methods. *Genomics Proteomics Bioinformatics* 2: 216-221

White,O., Dunning,T., Sutton,G., Adams,M., Venter,J.C. and Fields,C. (1993) A quality control algorithm for DNA sequencing projects. *Nucleic Acids Res.* 21: 3829–3838.

Xu,Y., Mural,R.J. and Uberbacher,E.C. (1994) Constructing gene models from accurately predicted exons: An application of dynamic programming. *Comput. Appl. Biosci.* 10: 613-623.

Yan,M., Lin,Z.-S. and Zhang,C.-T. (1998) A new Fourier transform approach for protein coding measure based on the format of the Z curve. *Bioinformatics* 14: 685-690.

5. CONCLUSÕES

As técnicas de identificação de genes independentes da espécie podem ser amplamente utilizadas de forma segura nos processos de anotação automática de genomas se forem levadas em consideração suas limitações de uso. São ferramentas que podem diminuir de forma considerável a quantidade de predições erradas.

As metodologias mais comuns para identificação de regiões codificantes independente de espécie, como a AMI, utilizam da mesma medida entrópica causada pela estrutura de códons para caracterizar as seqüências como codificantes ou não-codificantes. Esse poder de discriminação depende diretamente do tamanho da seqüência. Abaixo de 500 nucleotídeos, o poder discriminativo decai rapidamente, uma vez que em seqüências pequenas a quantidade de informação é muito pequena para ser detectada. Pseudogenes algumas vezes podem ser falsos positivos, pois a medida codificante permanece nas seqüências a não ser que elas sofram mutações aleatórias acima de 30% onde aproximadamente 60% dos códons sofram mutação em pelo menos um nucleotídeo. Porém, metade das seqüências codificantes já passa a ser considerada não codificantes com apenas 15% de mutações aleatórias.

O AMI é uma ferramenta complementar às metodologias que utilizam cadeias de Markov e Redes Neurais para anotação de genomas, uma vez que ela é capaz de eliminar maior parte dos falsos positivos evitando contaminações nos conjuntos de treinamento. Na utilização da AMI para certificar a anotação do genoma do arroz foi encontrado cerca de 10% de genes errados, tanto no GenBank quanto no TIGR. A AMI inserida no processo de anotação para validação de seqüências codificantes evitaria o grande número de genes mal preditos e seria muito útil nas primeiras etapas de anotação em genomas pouco conhecidos como o do *Moniliophtera perniciosa*.

A metodologia desenvolvida para identificação do quadro de leitura, FRAMER, soluciona essa deficiência das metodologias independentes da espécie. É útil no tratamento de organismos onde o nível de GC ou uso do códon não estão ainda identificados, não requer treinamento, banco de dados, nem informações prévias. Por ser também uma metodologia eficiente mesmo em seqüências muito curtas (~50 bp) ela se torna interessante para processamento de ESTs (~350 bp) de novas espécies ou tecidos com pouca quantidade de informações preexistentes, uma vez que até o Blastx e *ORFfinders* perde eficiência abaixo dos 200 bp.

O estudo e desenvolvimento dessas ferramentas de bioinformática forneceram para os projetos genoma desenvolvidos na Universidade Estadual de Santa Cruz suporte computacional para o processamento dos resultados de seqüenciamento derivados de novas espécies, desconhecidas no meio científico. Isso tornou possível a análise e descoberta de novos genes, o que não seria possível com as metodologias tradicionais.

6. REFERÊNCIAS COMPLEMENTARES

ABU-HANNA, A. **Review of "Machine Learning," Artificial Intelligence in Medicine**, Elsevier. V. 16, p. 201-204, 1999.

AGGARWAL, G.; RAMASWAMY, R. Ab initio gene identification: prokaryote genome annotation with GeneScan and GLIMMER. **J. Biosci**. V. 27 p. 7–14, 2002.

ALTSCHUL, S.; GISH, W.; MILLER, W.; MYERS, E.; LIPMAN, D. Basic local alignment search tool. **J. Mol. Biol.** V. 215, p. 403–410, 1990.

BELL, P.J.L.; HIGGINS, V.J.; DAWES, I.W.; BISSINGER, P.H. Tandemly repeated 147 bp elements cause structural and functional variation in divergent MAL promoters of Saccharomyces cerevisiae. **Yeast** . V. 13, p. 1135–1144, 1997.

BESEMER, J.; BORODOVSKY, M. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. **Nucleic Acids Research**. V. 33, p. 451-454, 2005.

BLUMENTHAL, T. Operons in eukaryotes. **Brief Funct Genomic Proteomic**. V. 3, p. 199-211, Nov. 2004.

BOHJANEN, P.R.; LIU, Y.; GARCIABLANCO, M.A. TAR RNA decoys inhibit Tat-activated HIV-1 transcription after preinitiation complex formation. **Nucleic Acids Res**. V. 25, p. 4481–4486, 1997.

BORODOVSKY, M.; MCININCH, J. GENEMARK: parallel gene recognition for both DNA strands. **Computers & Chemistry**. V. 17, p. 123-133, 1993.

BRUNAK, S.; ENGELBRECHT, J.; KNUDSEN, S. Prediction of human mRNA donor and acceptor sites from the DNA sequence. **J. Mol. Biol**. V. 220, p. 49-65, 1991.

BURATTI, E.; BARALLE, F.E. Influence of RNA Secondary Structure on the Pre-mRNA Splicing Process. **Molecular and Cellular Biology**. V. 24, n. 24, p. 10505-10514, Dez. 2004.

CARELS, N.; VIDAL, R.; MANSILLA, R.; FRÍAS, D. The mutual information theory for the certification of rice coding sequences. **FEBS Lett**. V. 568, p. 155-158, 2004.

CRUVEILLER, S.; JABBARI, K.; CLAY, O.; BERNARDI, G. Incorrectly predicted genes in rice? **Gene.** V. 333, p. 187-188, 2004.

DELCHER, A.L.; HARMON, D.; KASIF, S.; WHITE, O.; SALZBERG, S.L. Improved microbial gene identification with GLIMMER. **Nucleic Acids Res**. V. 27, p. 4636–4641, 1999.

FICKETT, J.W. Finding genes by computer: the state of the art. **Trends Genet**. V. 2, p. 316-320, Ago. 1996.

FICKETT, J.W.; HATZIGEORGIOU, AG. Eukaryotic promoter recognition. **Genome Res**. V. 9, p. 861-878, Set 1997.

FICKETT, J.W.; WASSERMAN, W.W. Discovery and modeling of transcriptional regulatory regions. **Current Opinion in Biotechnology**. V. 11, p. 19-24, 2000.

FRECH, K.; QUANDT, K.; WERNER, T. Muscle actin genes: a first step towards computational classification of tissue specific promoters. **In Silico Biol.** V. 1, p 5, 1998.

FUKUSHIMA, A.; IKEMURAB, T.; KINOUCHIE, M.; OSHIMAE, T.; KUDOD, Y.; MORIG, H.; KANAYA, S. Periodicity in prokaryotic and eukaryotic genomes identified by power spectrum analysis. **Gene.** V. 300, p. 203–211, 2002.

GELFAND, M.S.; MIRONOV, A.A.; PEVZNER, P.A. Gene recognition via spliced sequence alignment. **Proc Natl Acad Sci U S A**. V. 20, p.9061-9067, Ago. 1996.

GRIFFITH, A.J.F. et al. **An Introduction to Genetic Analysis**. W. H. Freeman**,** 2004. 800 p.

GROSSE, I.; BULDYREV, S.V.; STANLEY, H.E.; HOLSTE, D.; HERZEL, H. Average mutual information of coding and noncoding DNA. **Pac Symp Biocomput**. p. 614-23, 2000a.

GROSSE, I.; HERZEL, H.; BULDYREV, S.V.; STANLEY, H.E. Species independence of mutual information in coding and noncoding DNA. **Physical Review E.** V. 61, p. 5624-5633, Mai 2000b.

GUIGÓ, R. et al. An assessment of gene prediction accuracy in large DNA sequences. **Genome Res**. V. 10, p. 1631-1642, 2000

GUO, F.B.; OU, H.Y.; ZHANG, C.T. ZCURVE: a new system for recognizing proteincoding genes in bacterial and archaeal genomes. **Nucleic Acids Research**.V. 31, p. 1780-1789 , 2003.

HENDERSON, J.; SALZBERG, S.; FASMAN, K. Finding genes in human DNA with a hidden Markov model. **J. Comput. Biol.** V. 4, p. 127-141, 1997.

HERZEL, H.; GROßE, I. Correlations in DNA sequences: The role of protein coding segments. **Physical Review E**. V. 55, p. 800-810, Jan. 1997.

ISSAC, B.; SINGH, H.; KAUR, H.; RAGHAVA, G.P.S. Locating probable genes using fourier transform approach. **Bioinformatics**. V. 18, p. 196-197, 2002.

KANAYA, S.; YAMADA, Y.; KINOUCHI, M.; KUDO, Y.; IKEMURA, T., Codon Usage and tRNA Genes in Eukaryotes: Correlation of Codon Usage Diversity with Translation Efficiency and with CG-Dinucleotide Usage as Assessed by Multivariate Analysis. **Journal of Molecular Evolution**. V. 53, p. 290-298, Out. 2001.

KIM, J.; KLOOSTER, S.; SHAPIRO, D.J. Intrinsically bent DNA in a eukaryotic transcription factor recognition sequence potentiates transcription activation. **J Biol Chem.** V. 270, p. 1282–1288, 1995.

KLEFFE, J.; HERMANN, K.; VAHRSON, W.; WITTIG, B.; BRENDEL, V. Logitlinear models for the prediction of splice sites in plant pre-mRNA sequences. **Nucleic Acids Res.** V. 24, p. 4709-4718, 1996.

KNUDSEN, S. Promoter2.0: for the recognition of PolII promoter sequences. **Bioinformatics**. v. 5, p. 356-361, Mai. 1999.

LEWIS, S.; ASHBURNER, M; REESE, M.G. Annotating eukaryote genomes. **Current Opinion in Structural Biology**. V. 10, p. 349-354, 2000.

LI, W.; KANEKO, K. Long-range Correlation and Partial 1=f Spectrum in a Non-coding DNA Sequence. **Europhysics Letters**. V. 17, p. 655-660, 1992.

LI, W.; MARR, T.G.; KANEKO, K. Understanding long-range correlations in DNA sequences. **Physica D**. V. 75, p. 392–416, 1994.

LYNN, A. M.; JAIN, C. K.; KOSALAI, K.; BARMAN, P.; THAKUR, N.; BATRA, H.; BHATTACHARYA, A. An automated annotation tool for genomic DNA sequences using GeneScan and BLAST. J. **Genet**. V. 80, p. 9–16, 2001.

MANSILLA, R.; BUSH, E. Increase of complexity from classical Greek to Latin poetry. **Complex Systems**. V. 14, p. 1-28, 2002.

MARTINEZ-ABARCA, F.; TORO, N. Group II introns in the bacterial world. **Mol Microbiol**. V. 38, p. 917-26, Dez. 2000.

MATHE, C.; SAGOT, M.F.; SCHIEX, T.; ROUZE, P. Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res. V. 19, p. 03-17, Out. 2002.

MATTICK, J.S. Introns: evolution and function. **Current Opinion in Genetics and Development**. V. 4, p. 823-831, 1994.

OHLER, U. Promoter Prediction on a Genomic Scale---The Adh Experience. **Genome Res**. V. 10, p. 539-542, 2000.

OHLER, U.; NIEMANN, H.; LIAO, G.; RUBIN, G.M. Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition. **Bioinformatics**. V. 17, p. 199-206, 2001

PEDERSEN, A.G.; BALDI, P.; CHAUVIN, Y.; BRUNAK, S. The biology of eukaryotic promoter prediction - a review. **Comput. Chem**. V. 15, p. 191-207, 1999.

REESE, M.G. Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome. **Computers & Chemistry**. V. 26, p. 51-56, 2001.

REESE, M.G.; EECKMAN, F.H.; KULP, D.; HAUSSLER D. Improved splice site detection in Genie. In Istrail,S., Pevzner,P. and Waterman,M. (eds), First Annual International Conference on Computational Molecular Biology (RECOMB). **ACM Press**. p. 232-240. 1997.

ROGOZIN, I.B.; MILANESI, L. Analysis of donor splice signals in different organisms. **J. Mol. Evol**. V. 45, p. 50-59, 1997.

ROMBAUTS, S.; FLORQUIN, K.; LESCOT, M.; MARCHAL, K.; ROUZÉ, P.; PEER, Y.V. Computational approaches to identify promoters and cis-regulatory elements in plant genomes. **Plant Physiology**. V. 132, p. 1162-1176, Jul. 2003.

SALZBERG, S.; DELCHER, A.; FASMAN, K.; HENDERSON, J. A decision tree system for finding genes in DNA. **J. Comput. Biol**. V. 5, p. 667-680, 1998.

SALZBERG, S.L.; PERTEA, M.; DELCHER, A.L; GARDNER, M.J.; TETTELIN, H. Interpolated Markov Models for Eukaryotic Gene Finding. **Genomics**. V. 59, p. 24–31, 1999.

SAUL, A.; BATTISTUTTA, D. Codon usage in Plasmodium falciparum. **Mol. Biochem. Parasitol**. V. 27, p. 35–42, 1988.

SCARPARI, L.M.; MEINHARDT, L.W.; MAZZAFERA, P.; POMELLA, A.W.; SCHIAVINATO, M.A.; CASCARDO, J.C.; PEREIRA, G.A. Biochemical changes during the development of witches' broom: the most important disease of cocoa in Brazil caused by Crinipellis perniciosa. **J Exp Bot**. V. 413, p. 865-877, Mar 2005.

SHANNON, C.E. A mathematical theory of communication. **Bell Systems Technical Journal** V. 27, p. 379-423, 1948.

SMITH, T.F.; WATERMAN, M.S. Identification of common molecular subsequences. **JMol Biol**. V. 147, p.195-197. 1981

TIWARI, S.; RAMACHANDRAN, S.; BHATTACHARYA, A.; BHATTACHARYA, S.; RAMASWAMY, R. Prediction of probable genes by Fourier analysis of genomic sequences. **Comput. Appl. Biosci**. V. 13, p. 263-270, 1997.

TRIFONOV, EN.; SUSSMAN, J.L. The Pitch of Chromatin DNA is Reflected in its Nucleotide Sequence. PNAS. V. 77, p. 3816-3820, 1980.

UBERBACHER, U.C.; MURAL, R.J. Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. **Proc. Natl. Acad. Sci**. **USA**. V. 88, p. 11261-11265, 1991.

VAIDYANATHAN, P.P.; YOON, B. Digital filters for gene prediction applications. **Signals, Systems and Computers.** V. 1, p. 306- 310, Nov. 2002.

VINOGRADOV, A.E. Intron Length and Codon Usage. **Journal of Molecular Evolution**, v. 52, p. 2-5, Jan. 2001.

WANG, J.; ZHANG, C.T. Identification of protein-coding genes in the genome of Vibrio cholerae with more than 98% accuracy using occurrence frequencies of single nucleotides. **Eur J Biochem**. V. 268, p. 4261-4268, 2001.

WANG, Z.; YAZHU, C.; LI, Y. A brief review of computational gene prediction methods. **Geno. Prot. Bioinfo**. V. 2, p. 216-221, 2004.

WERNER, T. Models for prediction and recognition of eukaryotic promoters. **Mammalian Genome**. V. 10, p. 168-175, 1999.

WINGENDER, E.; CHEN, X.; HEHL, R.; KARAS, H.; LIEBICH, I.; MATYS, V.; MEINHARDT, T.; PRÜB, M.; REUTER, I.; SCHACHERER, F. TRANSFAC: an integrated system for gene expression regulation. **Nucl. Acids Res**. V. 28, p. 316-319, 2000.

YAN, M.; LIN, Z.S.; ZHANG, C.T. A new fourier transform approach for protein coding measure based on the format of the Z curve. **Bioinformatics**. V. 14, p. 685-690, 1998.

ZHANG, C.T.; WANG, J. Recognition of protein coding genes in the yeast genome at better than 95% accuracy based on the Z curve. **Nucl. Acids Res**. V. 28, p. 2804-2814, 2000.

ZHANG, C.T.; ZHANG, R.; OU, H.Y. The Z curve database: a graphic representation of genome sequences. **Bioinformatics**. V. 19, p. 593-599, 2003.

ZHANG, M.Q. Computational Prediction of Eukaryotic Protein-Coding Genes. **Nature Reviews**. V. 3 p. 698-710, 2002.

ZHANG, R.; ZHANG, C.T. Z curves, an intuitive tool for visualizing and analyzing the DNA sequences. **J. Biomol. Struct. Dyn**. V. 11, p. 767-782, 1994.

APÊNDICE A - The mutual information theory for the certification of rice coding sequences

Authors: Nicolas Carels[1], **Ramon Vidal[1]**, Ricardo Mansilla[2], Diego Frías[1]

Instituition: [1]Laboratorio de Bioinformatica, Universidade Estadual de Santa Cruz, Rodovia Ilheus/Itabuna km. 16, Ilheus, Bahia, Brazil. [2]Centro de Investigacıones Interdiciplinarias en Ciencias y Humanidades, Universidad Nacional Autonoma de Mexico, Ciudad Universitaria, Del. Coyoacan, Mexico, DF, México.

Publicado: FEBS Letters, 568 (2004) 155–158

FEBS 28490 FEBS Letters 568 (2004) 155–158

The mutual information theory for the certification of rice coding sequences

Nicolas Carels[a,*], Ramon Vidal[a], Ricardo Mansilla[b], Diego Frías[a]

[a]*Laboratório de Bioinformática, Universidade Estadual de Santa Cruz, Rodovia Ilhéus/Itabuna km. 16, Ilhéus, Bahia, Brazil*
[b]*Centro de Investigacíones Interdiciplinarias en Ciencias y Humanidades, Universidad Nacional Autónoma de México, Ciudad Universitaria, Del. Coyoacan, México, DF, Mexico*

Received 7 April 2004; accepted 13 May 2004

Available online 25 May 2004

Edited by Robert B. Russell

Abstract We report here the use of the mutual information theory for the certification of annotated rice coding sequences of both GenBank and TIGR databases. Considering coding sequences larger than 600 bp, we successfully screened out genes with aberrant compositional features. We found that they represent about 10% of both datasets after cleaning for gene redundancy. Most of the rejected accessions showed a different trend in GC3% vs GC2% plot compared to the set of accessions that have been published in international journals. This suggests the existence of a bias in the pattern recognition algorithms used by gene prediction programs.

Keywords: Rice; AMI; GenBank; TIGR; Bimodality

1. Introduction

The conclusions drawn by bioinformatics rely on data quality. However, the joint development of sequencing techniques and automated annotation is responsible for the vast majority of gene annotations by in silico techniques. A consequence of this is the apparition of a substantial number of annotations corresponding to uncharacterized proteins in the genomic databases. Such proteins are considered as ‘‘hypothetical’’, ‘‘similar’’ or ‘‘putative’’ when they are homologous to accessions with unknown function. When there is no homolog, the term used for classification is ‘‘unique’’. Protein families with such *fuzzy* annotations are natural targets for functional genomics. However, *fuzzy* annotations may represent false gene predictions [7]. Moreover, the actual practice of protein annotation relies mainly on previous annotations. If certification steps are missed, it may lead to a ‘‘snow ball’’ effect.

Most gene prediction methods use hidden Markov models (HMMs) [1,2,9,13,14] or neural networks [13,15,17,18]. In optimal conditions, the precision of those algorithms was claimed to be in the range of 92–95%. However, the efficiency of these algorithms mainly depends on the quality of the ‘‘training dataset’’ used to recognize the patterns of interest. The training set can be (i) contaminated with erroneous sequences, (ii) redundant with the validation set, and (iii) not representative of the true population. In addition, the error of any pattern recognition method increases rapidly as the sequence length decreases because of the increasing lack of statistical significance. As a consequence, the risk to draw conclusions from false positives or true negatives is real (see [11], for a discussion of these methods). This risk particularly applies to large scale analyses where whole genomes are taken into consideration. Systematic bias in gene determination may introduce ‘original’ behaviors and lead to publication of wrong conclusions (see [4]).

Recently, a new kind of method for coding sequence (CDS) identification has been proposed [6,10,16]. Those methods detect short range spatial correlations between nucleotides specific to CDS due to their codon structure. Such methods do not rely on training or learning steps. The most outstanding methods of this class are the Sequence Fourier Spectrum [16] and the Average Mutual Information (AMI) [6].

Based on a statistically extremely robust gene sample, we found that the distribution of AMI-certified CDS confirms that the relationship GC3% vs GC2% must be considered monotonous [4].

Rice gene number has been claimed to be in the range of ~50 000, about 50% of which do not have homology with those of *Arabidopsis* [19]. The careful examination of the distribution of those predicted genes for GC3% vs GC2% by Cruveiller et al. [4] showed that it is bi-univocal, since it has two linear trends with different slopes intercepting at low GC3 and GC2 levels. This distribution differs essentially from those observed for other known eukaryote species. It would have promoted massive changes in transcription, transcriptional regulation, or translation since the dicot–monocot divergence, however, experimental evidence of such phenomenon was not found. In this article, using AMI, we showed by reference to a statistically representative set of experimental genes that one of the two trends of the plot of rice gene for GC3% vs GC2% is effectively formed by false positives. It is the first time that an automatic procedure is proposed to eliminate the major part of these false positives. In addition, it also allows to infer rice gene number with higher confidence and to solve an ongoing polemic regarding weather or not the distribution of *Gramineae* genes is bimodal [20].

2. Materials and methods

We retrieved complete nuclear CDSs of *Oryza sativa* from GeneBank and TIGR. The GenBank (release 137–15 August 2003) CDS

* Corresponding author. Fax: +55-73-680-5226.
E-mail address: carels@uesc.br (N. Carels).

0014-5793/$22.00 © 2004 Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies.
doi:10.1016/j.febslet.2004.05.026

(36 104) were retrieved, using the Infobiogen server (see http://www.infobiogen.fr) and the ACNUC/QUERY retrieval system [5] with the options: t = cds et no o = plastid et no o = mitochondrion et no k = partial et no k = est.

We used the bibliographical references reported under the field MEDLINE in the features to build a dataset of experimentally proven genes as follows: (i) the MEDLINE identification numbers were used to retrieve the abstracts of the corresponding genes from the NCBI server (PubMed), through a CGI interface (PERL); (ii) those abstracts were then screened to eliminate mitochondrial and chloroplast genes as well as (retro)transposons and references based on any kind of automatic in silico process. The remaining sample was 405 CDS.

To eliminate the redundancy from CDS samples, we looked for homology between sequences using BLASTN with the "-e" option equal to 0.0001. A cleaning procedure was then applied to the BLASTN file in order to eliminate the sequences implied in a homologous pair with the highest hit when it was above a given identity level. The identity level above which two sequences were considered redundant was set at 90% over 90% of the homologous regions with the shorter sequence of the pair.

The coding sequences of the protein genes were used to calculate their GC level in all three codon positions using the software CODONW [12].

In the case of TIGR, we downloaded the file "all.cds" from: ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_1.0/all_chrs. It contains 56 056 putative CDS identified among 358 Mb of non-overlapping sequences for all 12 chromosomes. We then eliminated the partial sequences and the sequences including the words: "tRNA", "transposon", "element", "plast" and "mitoch". The remaining file (52 309) CDS was further cleaned for redundancy as described above.

The histograms of gene distribution and the contour plots were obtained with Scilab-2.7 (http://scilabsoft.inria.fr/) with a class interval of 2% GC (except for the sample of published genes where class interval was set to 5% GC to have an image with better rendering). The orthogonal regression lines were calculated as described by Jolicoeur [8].

The AMI analyses were performed as described in [6] on sequences higher than 600 bp, because with such a criterion, the discrimination efficiency between coding and non-coding sequences is >98%.

We also established a statistic of the annotations. To do this, we considered basically two classes: (i) one that we called *fuzzy* with annotations such as: "hypothetical", "similar", "putative", "unknown" and "unnamed" CDS; (ii) the other that we called *consistent* because the annotations to which it refers was affirmative. We did this classification in order to find the proportion of these annotations between both CDS samples either certified or not by the AMI. To do this, we counted the lines of CDS definition containing these words using the function "egrep" in Linux in the CDS files either certified or not by the AMI.

3. Results

The non-redundant dataset of nuclear rice CDS (25 649) was found to be 71.04% of the complete set available from GenBank (36 104). The plot of gene with respect to GC2% and GC3% was found to be "bi-univocal" as described by Cruveiller et al. [4]. This was true for the total GenBank sample as well as for the sample cleaned for redundancy (Fig. 1A). In contrast, the experimentally proven published genes larger than 600 bp (405) display a univocal (monotonous) distribution of GC3% vs GC2% (Fig. 1B).

In the conditions of the gene sample of Fig. 1B we found, using the AMI, a prediction efficiency of 98% for true positives. With this error rate in mind, we analyzed the non-redundant dataset of nuclear rice CDS of GenBank and TIGR.

In GenBank, 17 069 (66.55%) non-redundant genes were larger than 600 bp. Among them, the putative "false positives" (2424) eliminated by the AMI are plotted in Fig. 1C and the putative "true positives" (14 645 = 86%) in Fig. 1D. We found

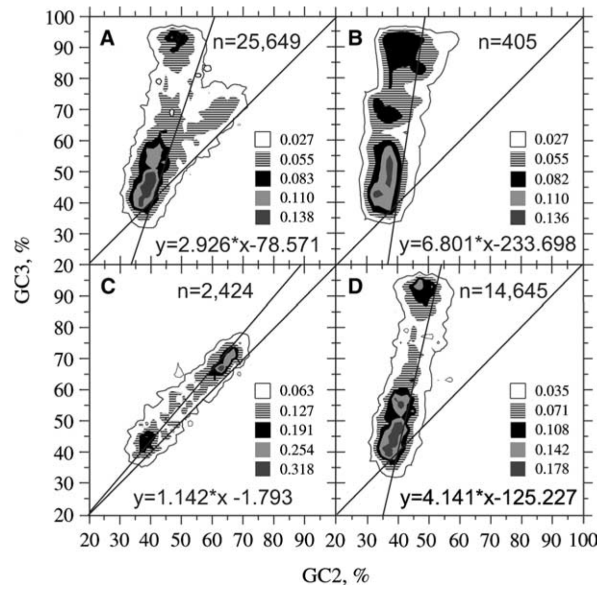


Fig. 1. Relationship between GC2 and GC3 in rice genes from: (A) the non-redundant dataset (25 649) of CDS >600 bp from GenBank, $r = 0.44$; (B) the non-redundant dataset (405) available from publications, $r = 0.35$; (C) the non-redundant dataset of B rejected by the AMI algorithm (2424), $r = 0.83$; and (D) the non-redundant dataset of B certified by the AMI algorithm (14 645), $r = 0.54$. The values for the contour lines indicate the relative density (%) of genes per unit area (%) of the plot. All correlation coefficients "r" were statistically significant at $P < 0.0001$. The different colored areas represent the different gene densities.

that Fig. 1D displays the same relationship as shown in Fig. 1B, with slight differences of the slopes of the regression lines probably due to the large sample size differences. In contrast, the slope of the regression line of Fig. 1C is close to the diagonal with the consequence that the GC2% of those "genes" should be approximately equal to GC3%. These sequences are classified as non-coding by the AMI and this is true with a rate of confidence of 98%.

In the GenBank dataset, we found 8353 (58%) and 1299 (54%) *fuzzy* annotations among the AMI-certified and AMI-rejected CDS, respectively.

In TIGR, the redundancy that we found using the criteria outlined above was about half (~14%) that of GenBank. We eliminated 7592 from a total of 52 309 CDS leaving 44 717 non-redundant genes. The redundancy in the TIGR dataset mainly affects the GC-poor genes (data not shown).

Among the 44 717 non-redundant genes, 27 979 (62.57%) were at least 600 bp; among which the AMI accepted 24 914 (89% = 24 914 of 27 979) and rejected 3065 (~11%) CDS. The contour plot of the 24 914 (Fig. 2B) and the CDS distribution for GC3% (Fig. 3A) of the TIGR dataset are similar to those of the 14 645 AMI-certified CDS from GenBank (Figs. 1D and 3B, respectively), leading to the confirmation that in GenBank as well as in TIGR the CDS distribution for GC3% is bimodal [3]. This bimodality of the CDS distribution tends to be hidden by the false positive genes as it was observed in GenBank and TIGR (Fig. 3).

Compared to the GenBank dataset (Fig. 1C), the profile of the rice CDS from TIGR rejected by the AMI (Fig. 2C) presents an interesting feature. The error of CDS prediction remains similar to that observed in GenBank (~10%). However,

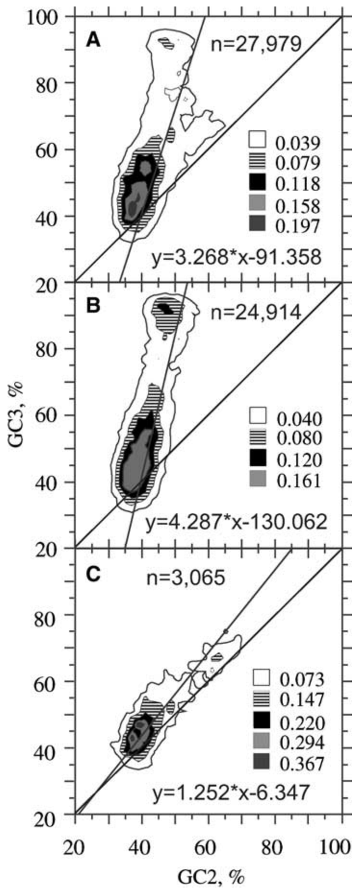


Fig. 2. Relationship between GC2 and GC3 in rice genes from: (A) the non-redundant dataset (27 979) of CDS >600 bp from TIGR, $r = 0.50$; (B) the non-redundant dataset of A certified by the AMI algorithm (24 914), $r = 0.58$; and (C) the non-redundant dataset of B rejected by the AMI algorithm (3065), $r = 0.78$. The values for the contour lines indicate the relative density (%) of genes per unit area (%) of the plot. All correlation coefficients "r" were statistically significant at $P < 0.0001$.

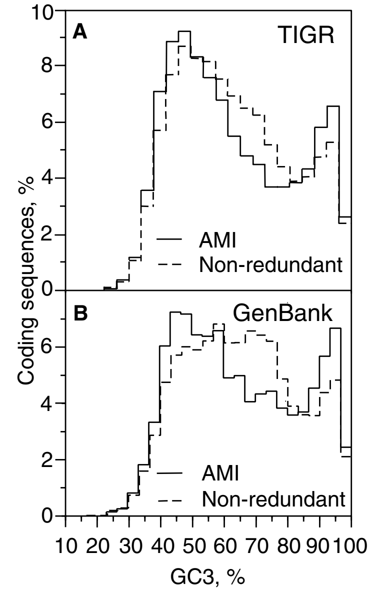


Fig. 3. Gene distribution with concern to GC3. The dotted line represents the non-redundant dataset and the solid line represents the AMI-certified CDS. (A) TIGR; (B) GenBank.

the shape of the contour plot of Fig. 2C deserves several comments. The proportion of miss-identified genes favors GC-poor genes. Moreover, the trend of the relationship between GC2% and GC3% appears similar to that of "true positives" (Fig. 2B) until 55% GC2–GC3. Then, the trend changes and becomes parallel to the diagonal. Compared to GenBank it is likely that the TIGR gene predictor implements a discrimination function sensible to the ratio GC3% vs GC2%. However, as far as we know for genomes such as *Arabidopsis*, where the GC3% range is about the same as that of GC2% (4% difference on the average), the prediction error (~10%, unpublished data) carried out by the gene predictor cannot be eliminated by methods other than those based on mutual information and Fourier transform.

In the TIGR dataset, we found 20 167 (81%) and 2754 (90%) *fuzzy* annotations among the AMI-certified and AMI-rejected CDS, respectively.

4. Discussion

The higher redundancy that we found in GenBank (30%) compared to TIGR (14%) is probably due to the fact that GenBank is a public repository. Therefore, numbers of genes can be entered independently as different alleles by different authors. Such a redundancy does not have biological meaning. In contrast, since only one group manages the TIGR database, it is expected that the 14% redundancy has a biological meaning and reflects paralogous genes in this genome. Those paralogous genes are probably more numerous, but the counting is relative and depends on the criteria that we applied to BLASTN for detection. Here, we were only interested in normalizing both datasets before any statistical treatment.

Based on coding sequences (CDS) larger than 600 bp, we found that the rate of wrongly annotated genes of rice in GenBank and TIGR was about ~10% of the non-redundant datasets. Therefore, the use of AMI appears to be a complementary method to HMM and neural networks, since it succeeded to eliminate genes that are obvious false positives by comparison to coding sequences that were decrypted by laboratory techniques. Those false positives display a random distribution along the line of identity between GC2% and GC3% (diagonal of Figs. 1C and 2C) and form a suspicious trend in the plot of rice genes for GC3% vs GC2% that is not present in the plot of the experimental gene set used as reference.

We found that in TIGR *fuzzy* annotations (~80%) are almost double that of GenBank (~50%). TIGR is a member of the IRGSP consortium (International Rice Genome Sequencing Project), which justifies that TIGR annotations are more complete than those of GenBank. It is likely that because of individual contribution much of the coding sequences to GenBank are simply not annotated. The proportion of *fuzzy* to *consistent* annotations in AMI-certified and AMI-rejected CDS appears to be roughly the same in GenBank and TIGR. It is interesting to conclude here that the process of annotation itself does not seem to be reliable, since a similar proportion of *fuzzy* and *consistent* annotations is found in AMI-certified as well as in AMI-rejected CDS (and this independently of the database). This suggests that, in addition to the *fuzzy* annotation, the *consistent* genes identified as false positives by the AMI are, indeed, miss-annotated. Our results suggest that AMI should be introduced in the IRGSP annotation protocol (see: http://rgp.dna.affrc.go.jp/genomicdata/AnnSystem.html) to take this problem into account.

The claim, on the basis of predicted genes, that *Gramineae* could have up to ~50% novel genes compared to *Arabidopsis* [19] remains to be confirmed. Considering only the 62.5% of non-redundant CDS of TIGR larger than 600 bp, we found that rice contains only 13.5% more CDS than *Arabidopsis*. Therefore, the ~26.5% genes that make up the difference between rice and *Arabidopsis*, under the Yu et al. [19] statement, most probably have CDS <600 bp. Since the error rate of gene prediction increases below this threshold, the consistency of the above claim remains to be confirmed.

Compositional bimodality of plant genes has been first reported by Carels and Bernardi [3]. In the case of *Gramineae* genes, the compositional heterogeneity is so large that it can be easily found just by looking at the profile of the gene distribution for GC3%. This evidence was denied by Meyers et al. [20] using maize ESTs. However, Meyers et al. [20] were wrong in their methodology. They did not find the bimodality, because they did not take EST redundancy and GC3% into consideration. With a coding sequence sample statistically extremely robust, Fig. 3 clearly shows that the rice coding sequence distribution for GC3% is, indeed, bimodal.

As a conclusion, the mis-understandings that occurred regarding: (i) the compositional bimodality of *Gramineae* genes, (ii) the bi-univocal gene distribution for GC3% vs GC2%, and (iii) the gene number in the rice genome are three different consequences of a same problem: mis-annotation. The certification of coding sequences of rice using AMI is a clear improvement to the present day knowledge about rice gene annotation. Preliminary results of coding sequences certification by AMI suggest that the method will also be very useful for the annotation of other genomes such as *Arabidopsis*, for instance.

Acknowledgements: We thank Carter Miller for the manuscript revision. This research was supported by the Brazilian CNPq agency providing researcher fellowship to N. Carels and student fellowship to R. Vidal.

References

[1] Besemer, J., Lomsadze, A. and Borodovsky, M. (2001) GeneMarkS: a self-training method for predicition of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res. 29, 2607–2618.

[2] Borodovsky, M. and McIninch, J.D. (1993) GeneMark: parallel gene recognition for both DNA strands. Comp. Chem. 17, 123–133.

[3] Carels, N. and Bernardi, G. (2000) Two classes of genes in plants. Genetics 154, 1819–1825.

[4] Cruveiller, S., Jabbari, K., Clay, O. and Bernardi, G. (2003) Compositional features of eukaryotic genomes for checking predicted genes. Brief Bioinform. 4, 43–52.

[5] Gouy, M., Gautier, C., Attimonelli, N., Lanave, C. and Di Paola, G. (1985) CABIOS 1, 167–172.

[6] Grosse, I., Herzel, H., Buldyrev, S.V. and Stanley, H.E. (2000) Species independence of mutual information in coding and noncoding DNA. Phys. Rev. E 61, 5624–5629.

[7] Iliopoulos, I., Tsoka, S., Andrade, M.A., Janssen, P., Audit, B., Tramontano, A., Valencia, A., Leroy, C., Sander, C. and Ouzounis, C.A. (2000) Genome sequences and great expectations. Genome Biol. 2, 1.1–1.3.

[8] Jolicoeur, P. (1963) The generalization of the allometry equation. Biometrics 19, 497–499.

[9] Kulp, D., Haussler, D., Reese, M.G. and Eeckman, F.H. (1996) A generalized hidden Markov model for the recognition of human genes in DNA. Proc. Int. Conf. Intelligent. Syst. Mol. Biol. 4, 134–142.

[10] Li, W., Marr, T.G. and Kaneko, K. (1994) Understanding long-range correlation in DNA sequences. Physica D 75, 392–416.

[11] Mathé, C., Sagot, M.F., Schiex, Th. and Rouzé, P. (2002) Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res. 30, 4103–4117.

[12] Peden, J. (1999). Analysis of codon usage, Ph.D. Thesis, University of Nottingham, UK. Available from <http://www.molbiol.ox.ac.uk/cu/>.

[13] Reese, M.G., Kulp, D., Tammana, H. and Haussler, D. (2000) Genie-Gene finding in *Drosophila melanogaster*. Genome Res. 10, 529–538.

[14] Salzberg, S.L., Delcher, A.L., Kasif, S. and White, O. (1998) Microbial gene identification using interpolated Markov models. Nucleic Acids Res. 26, 544–548.

[15] Sherriff, A. and Ott, J. (2001) Applications of neural networks for gene finding. Adv. Genet. 42, 287–297.

[16] Tiwary, S., Ramchandran, S., Bhattacharya, A., Bhattacharya, S. and Ramaswamy, R. (1997) Prediction of probably genes by Fourier analysis of genomic sequences. CABIOS 13, 263–270.

[17] Uberbacher, E.C. and Mural, R.J. (1991) Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. Proc. Natl. Acad. Sci. USA 88, 11261–11265.

[18] Xu, Y., Mural, R.J., Einstein, J.R., Shah, M.B. and Uberbacher, E.C. (1996) Grail: a multiagent neural network system for gene identification. Proc. IEEE 84, 1544–1551.

[19] Yu, J., Hu, S., Wang, J., Wong, G.K.-S., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y. and Zhang, X., et al. (2002) A draft sequence of the rice genome *Oryza sativa* L. ssp. *indica*). Science 296, 79–92.

[20] Meyers, B.C., Tingey, S.V. and Morgante, M. (2001) Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res 11, 1660–1676.

APÊNDICE B - Finding gene promoters in the genome of the fungus Crinipellis perniciosa using feed-forward neural networks

Authors: Diego Frias, **Ramon Vidal**, Julio Cascardo

Instituition: Laboratorio de Bioinformatica, Universidade Estadual de Santa Cruz, Rodovia Ilheus/Itabuna km. 16, Ilheus, Bahia, Brazil.

Publicado: Machine Learning for Signal Processing, 2004. Proceedings of the 2004 14th IEEE Signal Processing Society Workshop 29 Sept. - 1 Oct. 2004 Page(s):423 - 432 Digital Object Identifier 10.1109/MLSP.2004.1423003.

FINDING GENE PROMOTERS IN THE GENOME OF THE FUNGUS CRINIPELLIS PERNICIOSA USING FEED-FORWARD NEURAL NETWORKS

Diego Frías, Ramon Vidal and Júlio C.M. Cascardo

Bioinformatic Lab, Universidade Estadual de Santa Cruz, Ilhéus, Bahia, Brazil E-mail: diego@labbi.uesc.br

Abstract. The detection and structural characterization of genes in genome projects requires sophisticated automatic tools, most of then based on machine learning techniques. While functional genomics looks after the composition and function of the proteins codified by the genes, geneticists are more interested in investigating the mechanism, which regulates the expression of the genes. In particular, the study of the promoters is of crucial importance for understanding the responses to biological and environmental stimuli. In this article we address the use of neural networks for promoter recognition in the genome of the fungus Crinipellis perniciosa, an aggressive phytopathogen of the cacao tree, which is being sequenced by a Brazilian Consortium. A divide and conquer strategy was used for the solution of the complex problem of localizing and characterizing the gene promoters. The division of the problem is based on the localization of an internal structure called TATA-box, considered as one of the promoter's signal. With that purpose we trained a feed-forward neural network using patterns found in other species, due to absence of validated data for the fungus under study. A new approach for feature extraction, based on local compositional measures, is described. Currently, biological studies are being carried out for the experimental validation of the predictions of the neural network.

INTRODUCTION

High throughput techniques for DNA and RNA sequencing have driven the necessity of development of computational tools for sequence analysis. In particular, the detection and classification of genes in genome projects is a complex task that nowadays is implemented by sophisticated automatic annotation frameworks based mainly on probabilistic, like Markovian models [1, 2, 3], and machine learning techniques [4, 5]. Genes are distributed some how uniformly along each strand of the chromosomal sequence and are mostly constituted by a promoter region followed by a coding region [6]. The promoter region is responsible for attracting several macromolecules that, after binding to it, participate in the transcription process. The transcription process generates the messenger RNA (mRNA) that is a copy of the coding region of the gene. The mRNA suffers splicing (in eukaryotes) and, after translation, generates an amino acid chain. Post-translational processes occur and the nascent peptide chain becomes a functional protein. The level of gene expression, and consequently, the rate of production of a protein in the cell, in a given state, is strongly determined by the activity of the promoter of the corresponding gene under specific conditions [7]. For this reason, the characterization of the promoters of an organism, such as a pathogen, could help to understand its interaction with the host [9]. For this purpose, we explored several computational techniques for finding promoters in the unpublished DNA sequences of the fungus Crinipellis perniciosa (CP), an extremely aggressive pathogen of the Theobroma cacao, which is being sequenced by the Bahia State Genome Consortium (http://www.lge.ibi.unicamp.br/vassoura/). In this article we show the results obtained using neural networks for this purpose. The text is organized as follows: In section 2 we describe the pattern recognition problem showing its complexity and a simplified approach. The process of extraction of the features to be used as input of the neural network is described in section 3. In section 4 we describe the architecture of the neural networks and the training process based on a set of experimentally validated promoters of other eukaryotes that does not contain any promoters of the fungus under investigation. In section 5 we show the performance of a set of prototype neural networks having different number of neurons in the intermediate layer. Also in this section we show the output of the best configuration when the network screens a real CP DNA sequence containing a promoter. In section 6 we extract some conclusions of the study.

STATEMENT OF THE PROBLEM

The problem of finding promoter regions in eukaryotic DNA sequences is notoriously complex [10, 11]. In order to understand its complexity let us describe the structure of eukaryotic genes. Figure 1 shows the coding and promoter region of a generic gene separated by a non-coding region of variable length called 5' un-translated region (5'UTR). Such region, is flanked at the left by the Transcription Start Site (TSS), and at the right by the Translation Start Codon (TSC). The TSS determines the beginning of the mRNA resulting from the DNA transcription process, while the TSC determines the beginning of the protein obtained by translation of the mRNA code by the ribosomes. The promoter region is usually considered as formed by two contiguous sub-regions, the core and upstream promoter as shown in figure

1. The core promoter contains the TSS and other important sub-structures like the TATA-box, recognized by the TFIID protein and the Pol-III protein complex, respectively, which play a determinant role in the gene expression. The upstream promoter regions comprise the cis-elements, DNA sequences recognized by proteins called transcription factors.

It is now well known that the transcription factors, which bind at those specific DNA-substructures, may enhance or inhibit the transcription of the corresponding gene and play an essential role in the response of the cell to biological and environmental stresses. In order to delimitate the promoter regions in DNA sequences we need, at least, to find the left and right ends of this DNA structure, that is, to detect the outermost cis-element on the left and the TSS on the right. To author knowledge, there is not a computational method that could reliably detect the TSS, which defines the right end of the promoter, neither the cis-elements present in the upstream promoter. This is due to the following features of such important DNA substructures: (1) Both TSS and cis-elements are too small, containing only few bases, (2) the base-composition of such substructures is not conserved, that is, varies strongly from gene to gene within a specie and differs from specie to specie for homologous gene, (3) The sizes of the cis-elements and TSS as well as the distances from the TSC vary between different genes, which unable us to use searching algorithms based on pre-defined positions, that might be desirable.

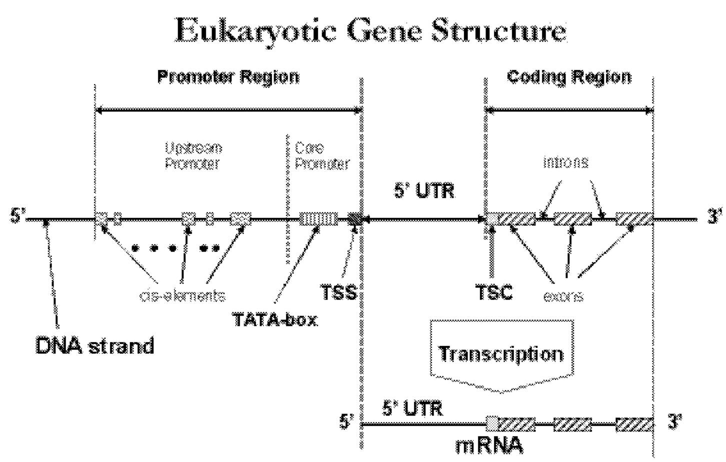


Figure 1. Illustration of the structure of genes in eukaryotes showing the promoter and coding regions and the transcription product.

Let's now turn to a practical challenge in using a machine learning approach for developing a computational tool for promoter-finding in the CP genomic data. Most machine learning algorithms must be trained with a statistically representative sample of data representing all the sought patterns [12]. In our case we are manipulating raw data, that means, there is not a set of certified CP promoters and non-promoter sequences that could be used for training the neural network. Under these conditions, we decided to use data available in the public Eukaryotic Promoter Database

TABLE 1: TATA-BOX SEQUENCES OF SEVERAL GENES FOUND IN THE EUKARYOTIC PROMOTER DATABASE

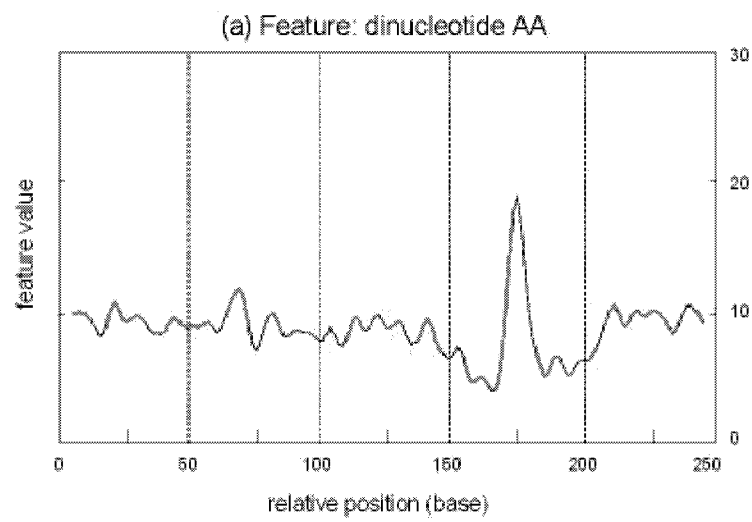
Specie	Seq ID	TATA-box	length
Zea Mays	EP07003	TATATAAATA	10
Drosophila melanogaster	EP23022	TATAAATA	8
Homo sapiens	EP11070	TATATAAAA	9
Bos taurus	EP07103	TATAAA	6
Xenopus laevis	EP14029	TATATAAA	8
Arabidopsis thaliana	EP24016	TATAAA	6
Consensus	Consensus	TATA T/A AAA T/A A	10

(http://www.epd.isb-sib.ch/) that contains 175 experimentally validated promoters of other species, including plants, vertebrates and homo sapiens. We knew that doing that choice implicitly we will require from the neural networks a significant degree of generalization that must reflects the pattern diversity present in the multi-specie training sample, associated to the species differentiation. Therefore, in this work we addressed the study of the capability of a neural network trained in a multi-species context for discriminating promoter from non-promoter regions in the CP genome.

Preliminary tests for TSS and cis-elements simultaneous identification showed that more robust techniques must be applied for solving the problem. Under this condition we turned for a three-stage divide and conquer strategy, with a first stage consisting in finding an inner reference substructure to be used: (1) as indicator of the presence of a hypothetical promoter region, and (2) for the decomposition of the DNA sequence into two subsequences, each one expected to contain a known subset of promoter substructures. By choosing the TATA-box as reference substructure the sequence lying on the left size must contain the cis-elements (upstream promoter), while the sequence on the right must contain the TSS (See figure 1). Specific techniques must be further applied for finding such well-differentiated substructures at each one of the two resulting sequences. In this paper we addressed the first step of the three-stage solution procedure for this complex problem, whose solution helped in obtaining a better knowledge of the promoter pattern specificities and allowed for designing approaches for the future solution of the remaining two sub-problems. The first problem can be stated as: To find a substructure s of length l into an input nucleotide sequence S of variable length $L \geq l$. In order to solve this problem within the machine learning approach we need a representative set of the sought substructure sequences $T = \{t_1, t_2, \dots\}$ of length l as well as an equivalent set $F = \{f_1, f_2, \dots\}$ with sequences of the same length but found in other DNA regions. In table 1 we show a subset of the sought TATA-box patterns extracted from the EPD. It can be observed the variability of both the length and base-composition of the target structures. In the last row of the table we included a 10 base consensus sequence for the complete set of promoters in such database.

FEATURE EXTRACTION PROCEDURE

The features to be used as input of the neural network were defined, among other possibilities, in terms of the local composition of the DNA sequence. The local composition was measured in terms of the absolute frequency of combinations of two and three nucleotides within a base-to-base sliding window of constant width. The window width was chosen to be equal to the average length of the sought promoter structure, that is, 10 base-pair for TATA-boxes. From a total of 80 available local compositional measures corresponding to all the combinations of two and three nucleotides, we considered a small subset of 8 measures as primary features for TATA-box recognition. The selection was based on the relative response of the local compositional measure when the sliding window passes through known TATA-boxes of a representative set of certified promoters. In order to illustrate the observed responses, in Figure 2 we show an example displaying the profiles of three features computed along a 250 base pairs promoter. In this example the TATA-box is found between bases 170 and 180 in the promoter sequence. Observe that all features deflect significantly within the TATA-box region displaying a smoother behavior outside. The other five features behave the same way in any case and were not included due to space limitation. This approach differs essentially from usual approaches to the problem of recognition of small patterns in nucleotide sequences. Most of them are based in nucleotide identification using an orthogonal codification, which relates a 4 bits binary number to each nucleotide [14]. However, the nucleotide codification, extensively applied for sequence analysis in prokaryotes (very simple organism), fails when applied to eukaryotes [10]. That is explained by the higher complexity of the eukaryotic genomes with respect to prokaryotic, reflected in a wider variability of the structures and patterns found in the eukaryotic DNA sequences.



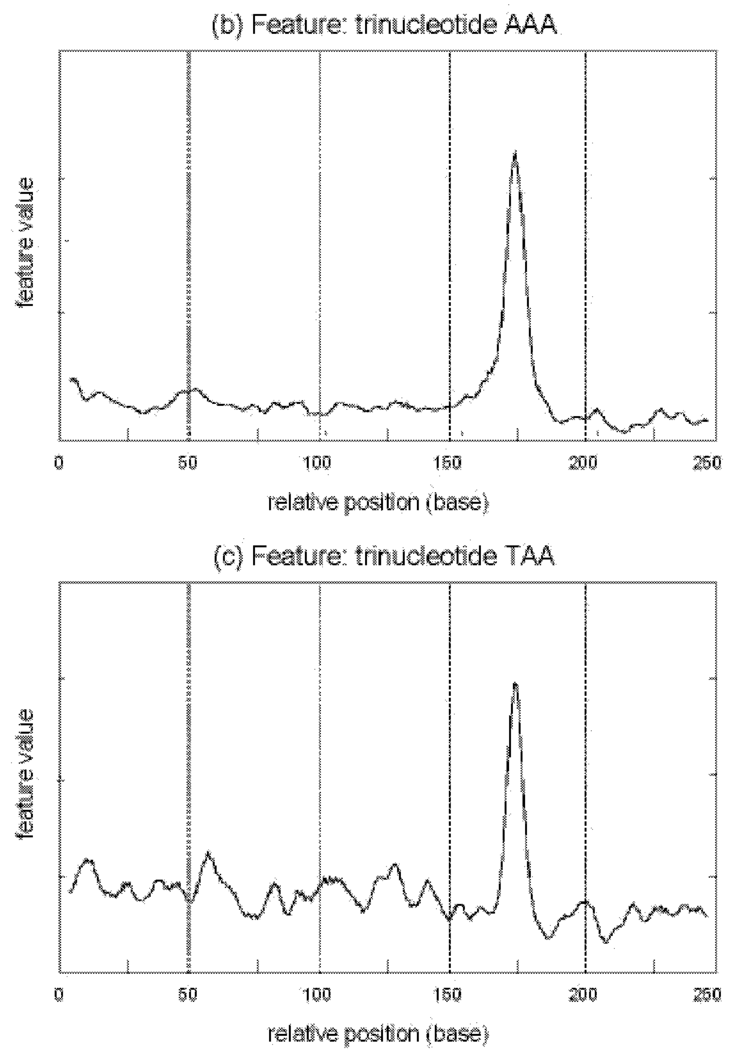


Figure 2.Profiles of three features computed with a 10 base width sliding window. The TATA-box in this example lies in the interval 170 -180, exactly where outstanding peaks occur in all cases.

NEURAL NETWORK CONFIGURATION AND TRAINING

Using SNNS [15] we built a feed-forward neural network with eight input neurons, one for each compositional feature, one hidden layer and one output neuron. The output was scaled in the interval $[0, 1]$ indicating the probability that the input pattern corresponds to a TATA-box. The number of hidden neurons was determined by testing different configurations. To measure the performance of each configuration we used the specificity S, and the sensitivity s, defined as

$$S = \left(1 - \frac{F_p}{N_F}\right) \times 100$$

$$s = \left(1 - \frac{F_n}{N_T}\right) \times 100$$

where F_p is the number of false positives, that is, the number of non-'TATA-box' patterns erroneously classified as TATA-boxes; F_n is the number of false negatives, given by the number of 'TATA-box' patterns erroneously classified; and N_T and N_F are the number of 'TATA-box' and non-'TATA-box' patterns in the validation data file, respectively. Extracting the TATA-box sequences from all the promoters we computed 175 'TATA-box' pattern-vectors containing the values of the 8 features. Also, cutting 10 bases from 200 arbitrarily chosen positions in the promoter sequences outside the TATA-box region, we constructed a set of non-'TATA-box' pattern-vectors. Three different sets of training and validation files were constructed from those patterns. Each set was constructed by collecting randomly the half of the TATA-box' and non-'TATA-box' patterns as training data, leaving the remaining patterns for validation purpose. Such files were formatted for supervised learning with SNNS, that is, in the first 8 columns we placed the pattern-vector and in the 9-th column a digit '0' for non-'TATA-box' pattern or digit '1' for 'TATA-box' pattern. Simultaneous back-propagation training and validation were performed for the three sets of data for a number of neurons increasing from 2 to 5 in the hidden layer. We recorded for each network configuration the smallest specificity and sensitivity values obtained in the validation step with the three sets of data. It was observed that the network with only two hidden neurons, as showed in Figure 3 had the best performance reaching a specificity of 99% and a sensitivity of 95%.

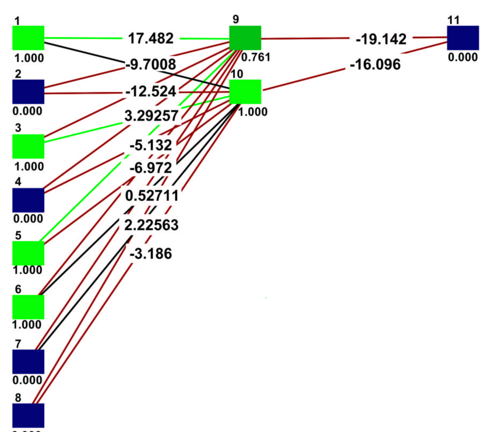


Figure 3.Best network for TATA-box recognition in eukaryotic DNA sequences

APPLICATION TO CP PROMOTER RECOGNITION

The neural network trained with TATA-box patterns present in a wide range of eukaryotes must be tested against the CP genome. In order to pick CP promoters, the DNA sequence of the fungus was compared with genes from a public database. Using the BLAST [13] code and the Genbank database (ftp://ftp.ncbi.nih.gov/blast/db) we found numerous matches with very high similarity between the CP sequence and the Genbank genes. That procedure, usually used for gene finding in bioinformatic applications [11], allowed the localization of several genes in the Crinipellis perniciosa DNA sequence. Twelve of such genes were arbitrarily chosen for experimental purposes. Based on biological statistics it was considered that core promoter structures like the TATA-box can be found in the first 600 bases preceding the Translation Start Codon (first ATG) of those genes (See figure 1). The neural network was then used for finding TATA-boxes over a set of twelve potential promoter sequences of 600 bases each. A ten bases width window was moved base-to-base along the whole sequences and the neural network outputs at each position was recorded. In Figure 4 we show the neural network output profile for one of such sequences. It can be observed several narrow small peaks distributed along the sequence and a unique wide and high peak, which indicates the existence of a putative TATA-box in such region. In fact, by simple inspection of the sequences inside the peak position (See the subsequence below the graph in capital letters) we found high similarity with known TATA-box patterns. This kind o output was observed for ten of the twelve potential promoter sequences showing a single well-defined region matching the TATA-box pattern at distances varying from 90 to 300 bases from the start codon, which is consistent with biological observations. The absence of putative TATA-boxes in the other two sequences it seems to be related to unusual promoter structure or degenerative mutations.

Taking into account that, according to a primary principle in bioinformatic applications, any result inferred by computational mean must be considered just as a conjecture from a biological point of view [1, 2, 3, 5], it was decided to investigate if such putative TATA-box regions found in the CP genome are really functional TATA-box structures of CP promoters. For this purpose the putative TATA-box sequences were extracted from the DNA of the fungus using proper primers and amplified by PCR [7]. The amplified DNA fragments were cloned on suitable vectors driven the expression of the reporter gene GUS [8]. The clones are being introduced in Crinipellis perniciosa protoplasts and plants, in order to analyze the functionality of the putative promoter, through the analysis of the GUS expression in the transgenic organisms. That is, GUS gene expression, signalized by a green coloration of the cell culture, indicates that the cloned fragment is a functional TATA-box. The results of the on-going biological test might elucidate whether or not multi-specie eukaryotic TATA-box patterns can be used for training machine learning tools to be used for finding TATA-box structures in the CP genome.

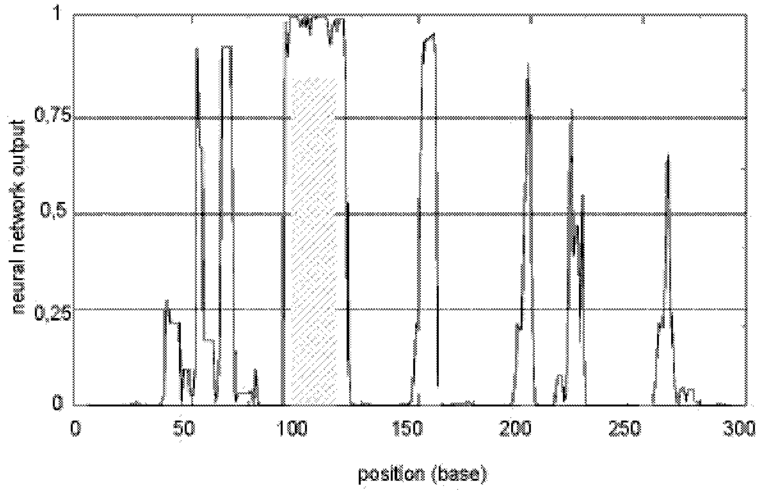


Figure 4.Output of the neural network for a CP promoter. The widest and highest peak matches the putative TATA-box region, represented in bold uppercase letters in the sequence below: "gtagatgtttcaccggaactgcaggtcatcacac caattagacctgaacagtattcttccaacccaagcaaccggtaagagaaataagggaattaccgATATA AAAAAATAgggaggccaaagtgtagtatcccatgtgagagagggaaggagaggaatgtttgacagt gttttggcggagaagttggcagctctcttcaatgtaccatttactcaaaggcttacacgaacatttgacaagg caagtacaactgcaattctgggtattcgagccactgaaactttagcagaaaggaaga"

CONCLUDING REMARKS

The recognition and characterization of eukaryotic promoters is a complex problem that can not be addressed simultaneously. The use of machine learning techniques like neural network seems to be a suggestive approach for solving smaller subproblems. In particular, the use of feed-forward multilayer perceptron network is a promissory approach for finding small structures like the TATA-box in DNA sequences, but its effectiveness will be always conditioned to the selection of the feature set.

The results of this work suggest that the use of features based on local compositional measures, like the dinucleotides and trinucleotides frequencies, instead of the usual nucleotide codification enhances the applicability of neural networks to the structural characterization of promoters and the study of the regulatory mechanism of genes.

Finally, considering the strong correlation between the function and the structure and composition of most biomolecules, it is highly expected the proposed approach to be validated by the on-going biological tests and the constructed neural network be capable to find TATA-box structures, not only in the Crinipellis perniciosa DNA sequences, but in any eukaryotic genome.

ACKNOWLEDGMENTS

This work was supported by FAPESB and CNPq grants.

References

[1] Borodovsky, M., and McIninch, J. D. (1993) "GeneMark: Parallel gene recognition for both DNA strands," Comp. Chem., 17, pp. 123-133

[2] Kulp D, Haussler D, Reese MG, Eeckman FH. A generalized hidden Markov model for the recognition of human genes in DNA. Proc Int Conf Intell Syst Mol Biol. 1996;4:134-42.

[3] Steven L. Salzberg, Arthur L. Delcher, Simon Kasif and Owen White. Microbial gene identification using interpolated Markov models. Nucleic Acids Research, 1998, Vol. 26, N.2

[4] Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information. S.M. Hebsgaard, P.G. Korning, N. Tolstrup, J. Engelbrecht, P. Rouze and S. Brunak, Nucleic Acids Research, 1996, Vol. 24, No. 17, 3439-3452.

[5] Reese, MG and Eeckman, FH, 1996. Splice Sites: A detailed neural network study", Proceedings of the 1996 Genome Mapping & Sequencing Meeting, Cold Spring Harbour, New York, arranged by D. Bentley, E. Green and P. Hieter.

[6] Voet D., Voet J. G. Biochemistry. Wiley, 1990, 852-888.

[7] Voet D., Voet J. G. Biochemistry. Wiley, 1990, 1032-1080.

[8] Jefferson RA, Kavanagh TA, Bevan MW.GUS fusions: beta-glucuronidase as a sensitive and versatile gene fusion marker in higher plants. EMBO J. 1987 Dec 20;6(13):3901-7.

[9] Gesteira Ada S, Micheli F, Ferreira CF, Cascardo JC. Isolation and purification of functional total RNA from different organs of cacao tree during its interaction with the pathogen Crinipellis perniciosa. Biotechniques. 2003 Sep;35(3):494-6, 498-500.

[10] J.W. Fickett et al, Eukaryotic promoter recognition, Genome Research 7(9) 861-878, 7997

[11] Uwe Ohler and Heinrich Niemann, Computational Identification and Analysis of Eukaryotic Promoters: New Algorithms on the Traces of Gene Regulation. Trends Genet. 2001 Feb; 17 (2):56-60

[12] Braga, A. Ludemir, T. e Carvalho, A. Redes Neurais Artificiais Teoria e Aplicações. LTC, 2000

[13] S.F. Altschul, et al., Basic local alignment search tool, J. Molec. Biol., 215(3):403-10, 1990.

[14] Qicheng Ma and Jason T. L. Wang. DNA Sequence Classification via an Expectation Maximization Algorithm and Neural Networks: A Case Study.

[15] A. Zell at al, SNNS Stuttgart Neural Network Simulator User Manual, Version 4.1, University of Stuttgart, Institute for Parallel and Distributed High Performance Systems.