Relatório de

TRABALHO FINAL DE CURSO

do Curso de
LICENCIATURA EM ENGENHARIA
INFORMÁTICA E DE COMPUTADORES (LEIC)

Ano Lectivo 2003 /2004

Departamento
de Engenharia
Informática

N.º da Proposta: 16

Título: Geração de código IUPAC

Professor Orientador:

Ana Teresa Freitas _____(assinatura)_____

Co-Orientador:

Arlindo Limede Oliveira _____(assinatura)_____

Alunos:

44060, Nuno Dias Mendes _____(assinatura)_____

48245, David Varela Nunes _____(assinatura)_____

Lista de Siglas

DNA (ou ADN) - ácido desoxirribonucleico

IUPAC - *International Union for Pure and Applied Chemistry*; designa, neste trabalho, o código de sequências degeneradas de bases nucleicas estabelecido por esta organização

RNA (ou ARN) - ácido ribonucleico

SMILE - *Structured Motif Inference and Evaluation*

Lista de Figuras

Lista de Tabelas

Agradecimentos

Gostaríamos de agradecer à nossa orientadora, Professora Ana Teresa Freitas, pela forma como sempre orientou o nosso trabalho, pela inestimável disponibilidade, pelo interesse demonstrado e pela ímpar capacidade de motivação.

Agradecemos também ao nosso co-orientador, Professor Arlindo Oliveira, pela sua disponibilidade e opinião crítica que sempre contribuiu para o enriquecimento do nosso trabalho.

À Susana Vinga, pelo importante contributo na análise estatística e pela sua diligente disponibilidade.

À Joana Fradinho, Luís Figueiredo e Petra Jelinkova, pela disponibilização de conjuntos de dados essenciais para a prossecução deste trabalho.

À Alexandra Carvalho, por ter tido a disponibilidade de ler a primeira versão deste relatório e contribuído com pertinentes sugestões.

Aos nossos colegas da sala 138, pela camaradagem e companhia nas longas horas passadas no INESC.

Aos colegas da sala 128, pela motivação e constante disponibilidade.

Resumo

A grande quantidade de dados disponíveis na área da Biologia requer métodos de extracção de conhecimento poderosos, sistemáticos e eficientes. Em particular, a identificação de sequências de nucleótidos, designadas de consensos, onde se ligam factores de transcrição responsáveis, em parte, pelo mecanismo de regulação génica é essencial para uma compreensão do papel de cada gene.

Actualmente, existem vários algoritmos cujo objectivo é encontrar sequências candidatas a sequências de consenso. Estes algoritmos podem ser divididos em duas classes: métodos restritivos e métodos prolíficos. Os métodos restritivos obtêm poucas respostas com elevada precisão, enquanto que os métodos prolíficos obtêm muitas respostas com baixa precisão, exigindo um esforço de pós-processamento.

Neste trabalho utilizamos o SMILE [1] que é um método prolífico para a extracção de sequências de consenso na região promotora dos genes. Face ao número significativo das respostas obtidas pretende-se, com este trabalho, avaliar possíveis soluções de pós-processamento, de modo a garantir a utilidade da informação extraída para os utilizadores do algoritmo. Assim, definiram-se dois grandes objectivos:

- A avaliação de métodos estatísticos que permitam aferir a significância biológica das respostas obtidas;
- A obtenção de uma descrição compacta da informação extraída.

Palavras-chave: Gene, regulação génica, região promotora, sequências degeneradas, IUPAC, avaliação estatística de biosequências, SMILE, extracção de motivos

1 Introdução

A crescente disponibilidade de dados sobre o genoma de vários organismos tem possibilitado o aparecimento de novos projectos na área da biologia. No entanto, a quantidade massiva de dados disponíveis torna impossível uma análise manual da informação, sendo cada vez mais premente a necessidade de um processamento automático dos dados obtidos.

Neste contexto, a identificação e modelação da região promotora dos genes assume um papel fundamental, pois as condições de activação e transcrição de um gene dependem das sequências de nucleótidos aí presentes.

As sequências de nucleótidos da região promotora que determinam a transcrição do gene são denominadas de promotores e são representadas por sequências de consenso ou, simplesmente, consensos. A literatura refere várias formas de descrição destes consensos, desde a utilização de uma matriz de pesos, a uma sequência de bases mais frequentes. Neste trabalho, utilizamos uma codificação IUPAC destas sequências, que corresponde a um conjunto de sequências degeneradas, em que cada símbolo representa um conjunto de nucleótidos possíveis para uma dada posição.

Actualmente, existem vários algoritmos cujo objectivo é encontrar sequências candidatas a sequências de consenso nas regiões promotoras. As sequências que são reportadas como possíveis consensos são designadas de motivos. Os algoritmos de pesquisa de motivos podem ser agrupados, de uma forma geral, em dois tipos: algoritmos restritivos e algoritmos prolíficos.

Os algoritmos restritivos concentram-se em encontrar poucos motivos, mas de relevância assinalável; enquanto que os algoritmos prolíficos preocupam-se em determinar todos os motivos que possam ser encontrados de acordo com determinados parâmetros de pesquisa.

Os algoritmos designados de prolíficos têm a vantagem de não fazerem suposições *a priori*. No entanto, a grande quantidade de motivos encontrados requer um esforço de pós-processamento significativo.

Este trabalho tem por objectivo o desenvolvimento de uma metodologia que permita o processamento automático dos motivos encontrados por um algoritmo que pode ser classificado de prolífico, o algoritmo SMILE [1].

No Capítulo 2 apresentamos uma descrição do problema em análise, evidenciando a motivação biológica e fazendo o seu enquadramento com os objectivos do trabalho.

No Capítulo 3 descrevemos o algoritmo SMILE no contexto do problema da extracção de motivos e apresentamos um gerador de dados sintéticos que será usado como ponto de partida do estudo da eficácia do algoritmo.

No Capítulo 4 discutimos a utilização de métodos estatísticos para aferir a significância biológica dos motivos extraídos pelo SMILE.

No Capítulo 5 apresentamos um método para geração de uma representação compacta de motivos extraídos, usando a codificação IUPAC.

Finalmente, no Capítulo 6, apresentamos vários resultados obtidos tanto com dados sintéticos como com dados reais, discutindo a aplicabilidade dos métodos estatísticos apresentados no Capítulo 4.

2 Descrição do Problema

A molécula de DNA é responsável pela manutenção da informação genética de cada indivíduo e é constituída por duas cadeias de nucleótidos dispostas de modo a formar uma dupla hélice [2] que se encontra representada na Figura 2.1. Cada nucleótido é formado por três elementos: um açúcar, um grupo fosfato e uma de quatro bases azotadas. As quatro bases são a Adenina (A), a Timina (T), a Guanina (G) e a Citosina (C) e podem ser vistas como os quatro elementos de um alfabeto.

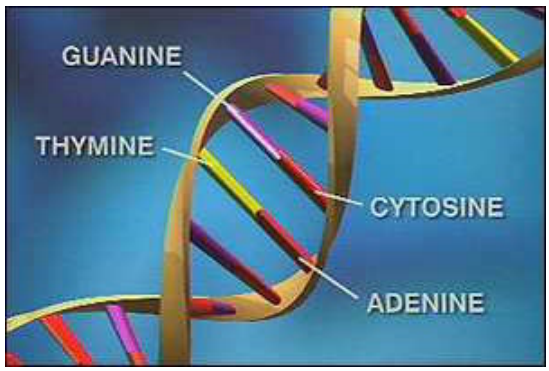


Figura 2.1 – Representação de cadeias de DNA em forma de dupla hélice

As duas cadeias de DNA são complementares no sentido em que as bases das duas cadeias estão emparelhadas. Assim, uma Adenina de uma cadeia emparelha com uma Timina da cadeia complementar e, do mesmo modo, uma Citosina emparelha com uma Guanina.

Os genes são segmentos de uma cadeia de DNA que controlam a síntese proteica e, consequentemente, toda a actividade celular. O genoma é o conjunto de genes de um indivíduo.

A expressão génica é o processo através do qual a informação contida nos genes dirige a síntese de proteínas. Este processo envolve dois passos: transcrição e tradução.

A transcrição consiste na transferência da informação contida no DNA para uma molécula intermediária de RNA designada de RNA-mensageiro ou mRNA. De seguida, o molde de mRNA é usado no passo de tradução onde irá dirigir a síntese proteica.

Este processo é semelhante em todos os organismos vivos embora existam diferenças assinaláveis entre organismos eucariotas e procariotas, isto é, entre organismos com e sem

núcleo individualizado, respectivamente. Em particular, o mecanismo é consideravelmente mais complexo em organismos eucariotas. A Figura 2.2 ilustra o processo de expressão génica em eucariotas (Figura 2.2-A) e procariotas (Figura 2.2-B).

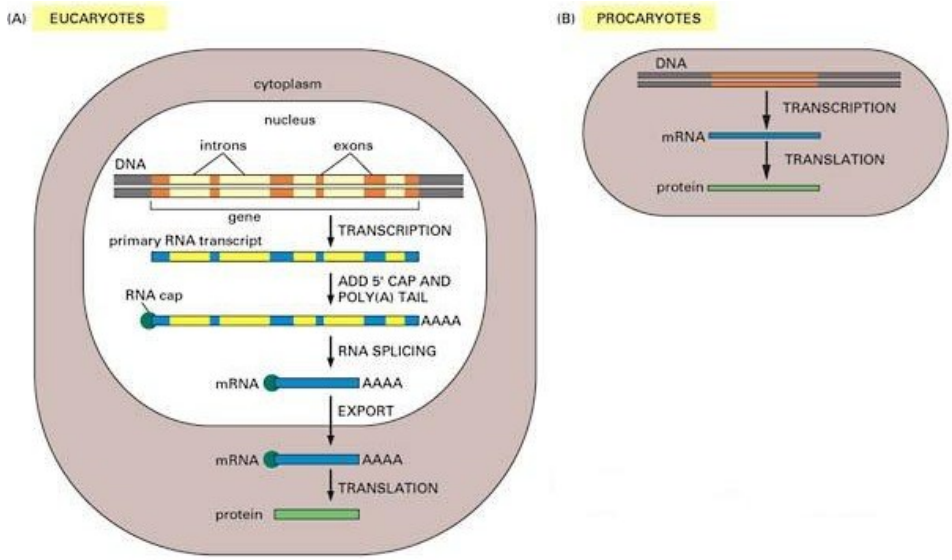


Figura 2.2 – Expressão génica em eucariotas (A) e procariotas (B).

Actualmente, o problema da identificação e modelação dos mecanismos que regulam a interacção entre grupos de genes é um dos problemas mais em foco na investigação em biologia computacional. Como já foi referido, a região promotora é parte essencial do mecanismo de regulação da transcrição génica.

A biologia oferece-nos alguns modelos da região promotora, evidenciando as diferenças entre organismos procariotas (Figura 2.3) e eucariotas (Figura 2.4).

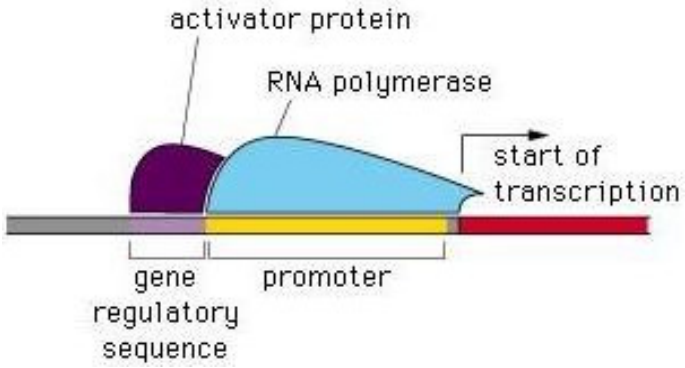


Figura 2.3 – Esquema da regulação génica ao nível da transcrição, em procariotas

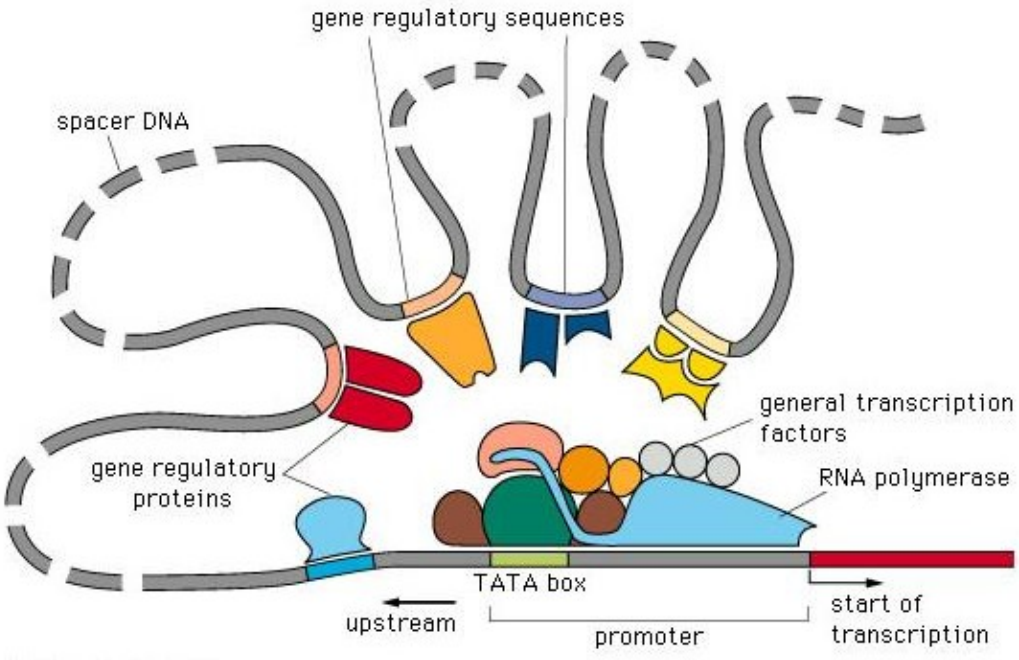


Figura 2.4 – Esquema da regulação génica ao nível da transcrição, em eucariotas

No mecanismo de transcrição génica intervém uma enzima denominada RNA-polimerase. Esta enzima tem como função a transcrição da informação contida no DNA para uma cadeia de RNA que irá posteriormente intervir no processo de síntese proteica.

Para que a RNA-polimerase inicie a transcrição de uma cadeia de DNA, geralmente um gene, é necessário que se cumpram algumas condições. Assim, entre outras, é necessário que

estejam presentes na região promotora motivos que a RNA-polimerase reconheça e aos quais se possa ligar para iniciar a transcrição da região a jusante (na literatura inglesa, estas regiões são denominadas de *binding sites*). As condições de activação desta enzima são mais complexas nos organismos eucariotas [3], uma vez que, neste caso, para além dos motivos referidos, existem ainda vários outros que são reconhecidos por factores de transcrição, isto é, proteínas, que se ligam sucessivamente à RNA-polimerase, estabelecendo um complexo mecanismo de regulação. Estes motivos podem estar a montante, a jusante, ou mesmo no interior do gene a ser transcrito e podem ainda exercer um efeito estimulador ou repressor da transcrição. Os locais onde existem motivos aos quais se ligam os diversos factores de transcrição são designados, genericamente, por regiões reguladoras.

É possível compreender agora a importância da necessidade de ser definido um modelo ou modelos para a região promotora, no âmbito do estudo do genoma.

Neste trabalho utilizamos o algoritmo SMILE para procurar motivos interessantes, isto é, biologicamente significativos, em cadeias de DNA pertencentes à região promotora dos genes. Como já foi referido, trata-se de um algoritmo prolífico, no sentido em que se preocupa em encontrar todas as sequências que cumpram os parâmetros especificados. Sendo assim, torna-se necessário um esforço significativo de pós-processamento para que possa ser obtida informação útil a partir dos dados gerados.

O pós-processamento dos motivos extraídos inclui uma análise estatística que culmina com a atribuição de um nível de significância a cada motivo. Deste modo, é possível definir um crivo que separa motivos considerados relevantes de motivos, à partida, menos interessantes.

Finalmente, e depois de ter sido encontrado um conjunto de motivos relevantes, será efectuada uma tradução para uma representação IUPAC. Esta representação permite compactar os dados, agrupando motivos muito semelhantes. Este passo é de importância essencial para uma utilização eficiente dos dados obtidos, uma vez que uma representação compacta facilita a sua utilização em pesquisas em bases de dados biológicas que contêm informação sobre consensos anteriormente identificados.

3 Extracção de motivos

3.1 Algoritmo SMILE

O SMILE é um método algorítmico de extracção de motivos de sequências biológicas. O algoritmo começa por processar um conjunto de dados de entrada composto por várias sequências não alinhadas de nucleótidos que se sabe conterem um ou mais *binding sites*. Estes dados são utilizados para calcular ocorrências de motivos usando como estrutura de dados base uma árvore de sufixos.

Um modelo é uma representação de subsequências que ocorrem nos dados de entrada satisfazendo, potencialmente, certas propriedades.

O algoritmo irá, assim, pesquisar modelos (simples ou estruturados) nos dados que lhe são fornecidos, segundo alguns parâmetros.

Um modelo simples diz respeito a uma subsequência contígua de nucleótidos enquanto que um modelo estruturado é uma composição de modelos simples.

Um modelo estruturado pode ser constituído por várias partes posicionadas ao longo da sequência. Cada parte é um modelo simples que pode ter um tamanho variável (entre *kmin* e *kmax*). O espaço entre cada parte consecutiva pode também ser parametrizado (*dmin* e *dmax*). Do mesmo modo, também é possível parametrizar o número de erros que se permite ocorrer no modelo e em cada uma das suas partes.

A Figura 3.1 apresenta um exemplo de uma especificação de um modelo estruturado em que se consideram duas partes. A primeira parte pode ter um tamanho entre 6 e 8 nucleótidos e são admitidos 2 erros. A segunda parte pode ocorrer entre 16 e 18 nucleótidos após a primeira parte e terá um tamanho igualmente entre 6 e 8 embora seja admitido apenas 1 erro.

A Figura 3.2 apresenta o esquema de outro modelo estruturado e respectiva concretização numa sequência. Neste exemplo consideram-se igualmente duas partes. A primeira com tamanho entre 6 e 10 nucleótidos e a segunda entre 6 e 8. As partes podem estar a uma distância que varia entre 10 e 15 nucleótidos e não são admitidos erros.

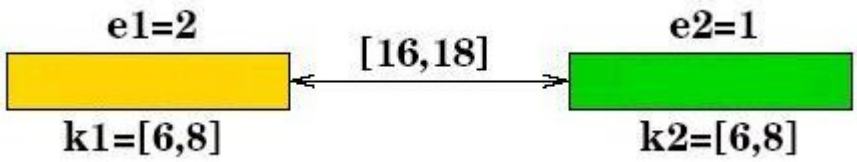


Figura 3.1 – Exemplo de uma especificação de um modelo estruturado

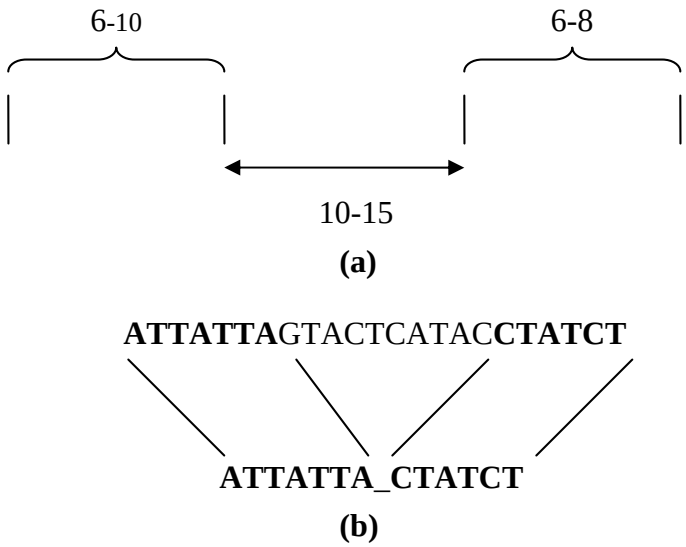


Figura 3.2 – (a) Exemplo de especificação; (b) Modelo retirado de uma sequência que cumpre a especificação.

Um modelo é considerado válido caso o número de sequências de entrada em que ocorre seja não inferior a um mínimo estabelecido, designado por quórum. Os modelos válidos são designados de motivos.

Para verificar o número de ocorrências é utilizada uma árvore de sufixos. Uma árvore de sufixos é uma estrutura de dados que representa todos os sufixos de uma cadeia de caracteres. As suas características são de grande utilidade para os algoritmos de emparelhamento de caracteres.

O SMILE processa conjuntos de dados, isto é, conjuntos de sequências de nucleótidos, utilizando, para esse efeito, uma árvores de sufixos generalizada. Esta estrutura de dados difere da árvore de sufixos simples no sentido em que agrupa várias cadeias de caracteres correspondentes às várias sequências de nucleótidos. Esta multiplicidade de sequências exige que a estrutura registe as sequências a que cada sufixo pertence. As árvores de sufixos fazem uso de um símbolo terminador ($) para assinalar o fim de cada sufixo.

A Figura 3.3 é um exemplo de uma árvore de sufixos generalizada.

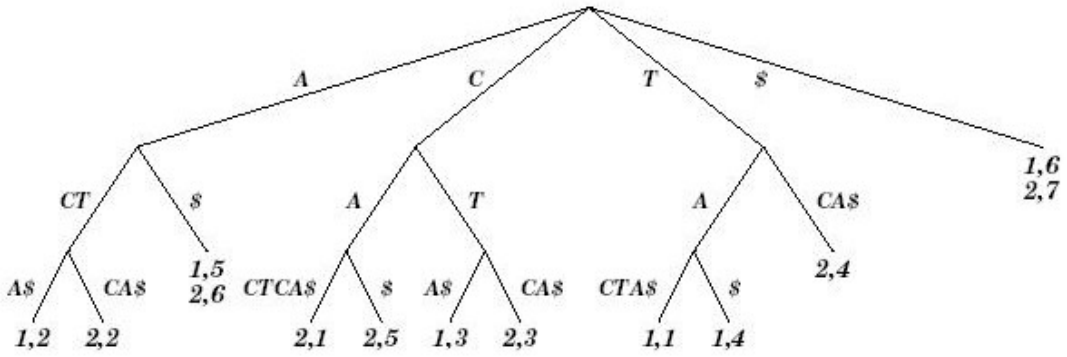


Figura 3.3 - Árvore de sufixos generalizada para as sequências TACTA$ (sequência 1) e CACTCA$ (sequência 2)

A extracção de motivos é feita através da travessia em profundidade-primeiro de uma árvore lexicográfica virtual, de modo a garantir que se explora o espaço de todos os modelos possíveis. Esta travessia da árvore lexicográfica é acompanhada da exploração da árvore de sufixos generalizada de forma a obter todos os sufixos a distância de *Hamming* não superior a e[1] do modelo em consideração. Assim, para cada nó da árvore virtual é verificado o quórum do modelo correspondente contabilizando, para tanto, o número de sequências diferentes que participam com sufixos a distância de *Hamming* não superior a *e* do modelo.

[1] O valor desta distância depende do número de erros admitidos na extracção de motivos e corresponde a um dos parâmetros do algoritmo.

A extracção de motivos estruturados é feita de forma semelhante, correspondendo a múltiplas procuras de modelos simples que constituem cada parte do modelo estruturado. A transição entre partes, ilustrada pela Figura 3.4, corresponde a um salto na árvore de sufixos com tantos níveis de profundidade adicionais quanto a distância especificada entre cada parte.

Uma descrição pormenorizada do algoritmo encontra-se em [4] e [5].

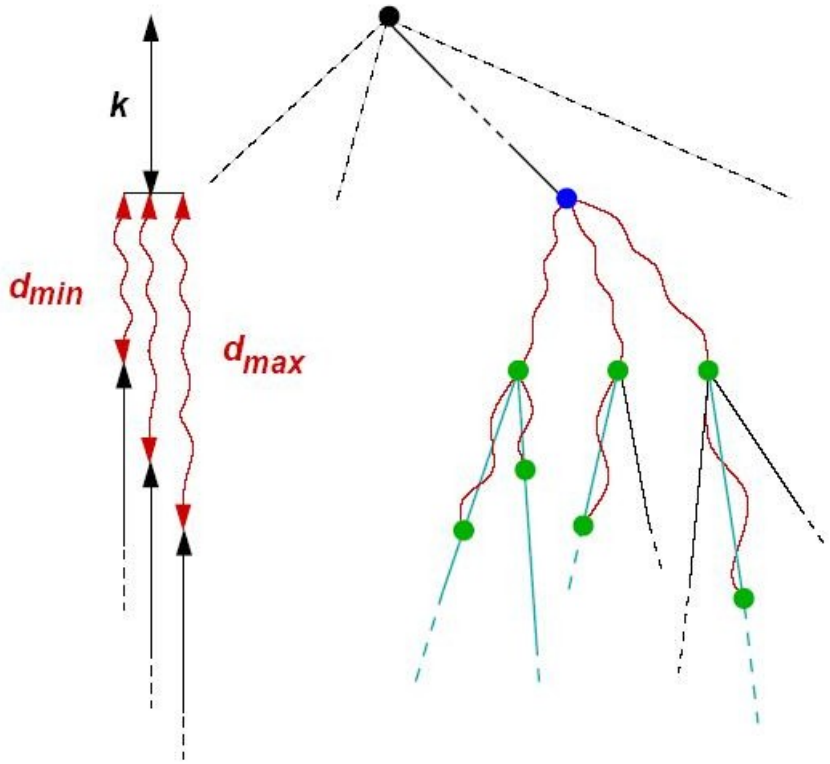


Figura 3.4 – Esquema da procura de motivos estruturados usando uma árvore de sufixos

Este algoritmo exacto gera todos os possíveis motivos dentro dos parâmetros especificados. Daqui resulta, geralmente, um número demasiado grande de motivos, sendo necessário discriminar quais os verdadeiros *binding sites*. Como método de avaliação, o SMILE utiliza um teste estatístico que é descrito na Secção 4.1.

3.2 Gerador de dados sintéticos

A análise da eficácia e eficiência do algoritmo de extracção de motivos exige a utilização de dados sobre os quais tenhamos, à partida, algum conhecimento. Para tal, foi desenvolvido um gerador de dados sintéticos.

O gerador de dados sintéticos permite, por um lado, evitar a necessidade de dispor de conjuntos de dados reais e, por outro, ter um maior controlo sobre os motivos que estão realmente presentes nos dados.

O gerador permite especificar o número e o tamanho das sequências do conjunto de dados, bem como os motivos a inserir. Os motivos são especificados pela sua sequência de nucleótidos, posição na sequência e frequência de ocorrência. Os motivos podem ser simples ou estruturados. A Figura 3.5 representa um exemplo de um ficheiro de especificação.

Frequência	*Posição*	*Sequência*	*Posição*	*Sequência*
0.2	09	TGCTTTTTAAT	28	TCGA
0.3	02	AGTACATCGA		
0.2	12	TGCGCA		

Figura 3.5 – Exemplo de um ficheiro de especificação para a geração de dados sintéticos

No exemplo apresentado são especificados três motivos (um motivo estruturado e dois motivos simples). A primeira coluna (*Frequência*) indica a percentagem de sequências em que o motivo deverá ocorrer. Em particular, para esta especificação, o primeiro motivo deverá ocorrer em 20% das sequências. As restantes colunas especificam a posição e a sequência de nucleótidos de cada parte do motivo. O número e o tamanho das sequências a gerar são especificados na linha de comandos do gerador.

O gerador verifica a especificação de modo a testar se é possível gerar um conjunto de dados com as características indicadas. Em particular, é necessário que, no caso das frequências especificadas somarem mais do que 1, alguns dos motivos sejam sobreponíveis uma vez que o gerador garante o cumprimento exacto da frequência de ocorrência especificada para cada motivo.

O processo de geração do conjunto de dados envolve duas fases. Numa primeira fase é gerado um conjunto de dados aleatórios em que a ocorrência de cada nucleótido é equiprovável. Para gerar aleatoriamente nucleótidos é usada uma variação do algoritmo *ran2* [6].

Numa segunda fase são inseridos os motivos especificados em tantas sequências quanto as necessárias para perfazer a frequência indicada. As sequências seleccionadas para a inserção de motivos são igualmente escolhidas de forma aleatória.

O recurso a dados sintéticos permite afastar a possibilidade de motivos bem classificados corresponderem a consensos ainda não identificados por via experimental, como pode acontecer quando se analisam conjuntos de dados reais. No entanto, é necessário termos presentes as limitações da actual abordagem. Em particular, sabemos que a geração equiprovável de nucleótidos está em desacordo com a estatística típica da região promotora, onde são mais comuns A's e T's do que G's e C's. Do mesmo modo, podem existir outras correlações de ordem superior, designadamente uma preferência por certos dímeros ou trímeros, em detrimento de outros. Esta limitação é ilustrada pelo facto de os resultados obtidos com dados sintéticos serem geralmente melhores do que os obtidos com dados reais. As causas e consequências desta observação serão discutidas nas secções seguintes.

4 Análise Estatística

4.1 Teste do χ^2

Como já foi referido, o SMILE, em geral, extrai um elevado número de modelos. Para isolar motivos relevantes é usado um teste estatístico de modo a atribuir uma significância estatística a cada modelo extraído.

O método de avaliação utilizado consiste em aplicar o teste do χ^2 com um grau de liberdade. O objectivo é determinar a probabilidade de obtermos motivos observados em sequências aleatórias que preservam os *k-mers*[2] das sequências originais. Assim, pretende-se perceber se a ocorrência do motivo se deve à frequência de *k-mers* presentes na sequência aleatória, tendo, portanto, uma baixa significância estatística, ou se é verificada a situação oposta, ou seja, elevada significância estatística. Estes últimos são considerados pelo algoritmo como os que apresentam maior potencial de serem verdadeiros *binding sites*. Para tal, o SMILE permite a realização de dois tipos alternativos de teste, um quanto ao número de sequências em que o motivo ocorre (quórum) e outro em relação à abundância do motivo no conjunto de dados.

Para aplicar o teste do χ^2 procedemos à construção de uma tabela de contingência com 4 entradas (Tabela 4.1).

	Dados Originais	Dados Aleatórios
Presença	P_{orig}	P_{random}
Ausência	A_{orig}	A_{random}

Tabela 4.1 – Tabela de contingência para o cálculo do χ^2

[2] Sequências de nucleótidos de tamanho *k*

As colunas da tabela referem-se às duas amostras em análise, ou seja, o conjunto de dados original e as sequências aleatórias geradas. As linhas da tabela referem-se às duas classes de contagens. Caso consideremos o teste quanto ao quórum, a classe Presença diz respeito ao número de sequências em que o motivo ocorre e a classe Ausência refere-se ao número de sequências em que o motivo não ocorre. Caso se trate do teste relativo à abundância, a primeira classe refere-se ao número de ocorrências do motivo no conjunto de dados e a segunda ao número de não ocorrências do motivo, em relação ao número de oportunidades de ocorrência. Considera-se como número de oportunidades de ocorrência o tamanho do conjunto de dados. Esta tabela será usada agora para fazer um teste de homogeneidade de proporções.

Para calcular a significância estatística basta usar a expressão do χ^2 correspondente à equação (4.1), onde O_{ij} diz respeito ao valor observado para a entrada *(i,j)* da tabela, e E_{ij} se refere ao correspondente valor esperado.

$$\chi^2 = \sum_{i,j=1,2} \frac{(E_{ij} - O_{ij})^2}{E_{ij}} \tag{4.1}$$

O valor esperado para cada entrada da tabela é dado pela expressão (4.2), ou seja, o valor esperado para a entrada *(i,j)* da tabela é igual ao produto do total da linha *i* pelo total da linha *j*, divido pela soma de todos os elementos da tabela.

$$E_{ij} = \frac{\sum_{k=1,2} O_{ik} \sum_{k=1,2} O_{kj}}{\sum_{k,l=1,2} O_{kl}} \tag{4.2}$$

Nas condições desta tabela de contingência é possível utilizar uma forma expedita de cálculo, que consiste na aplicação da expressão (4.3).

$$\chi^2 = \frac{(P_{orig}A_{random} - P_{random}A_{orig})^2 (P_{orig} + P_{random} + A_{orig} + A_{random})}{(P_{orig} + P_{random})(A_{orig} + A_{random})(P_{random} + A_{random})(P_{orig} + A_{orig})} \tag{4.3}$$

A hipótese nula que consideramos neste teste estatístico refere-se à possibilidade de os motivos terem sido gerados por sequências aleatórias sob certas restrições. Isto é, assumimos que as contagens obtidas para os motivos nas sequências originais são semelhantes às que obteríamos em sequências aleatórias com as mesmas frequências de *k-mers*. Valores altos de significância estatística implicam a rejeição da hipótese nula.

Um importante problema prende-se com a validade da aplicação do teste do χ^2. O SMILE, em particular, apresenta para cada motivo uma medida da sua significância estatística. No entanto, o verdadeiro interesse reside na determinação da significância biológica dos motivos encontrados. O que está em causa é perceber se um motivo é um potencial *binding site*. É necessário testar este método em dados reais com consensos conhecidos para se perceber se a significância estatística apresenta uma elevada correlação com a significância biológica.

No Capítulo 6 apresentamos uma discussão acerca da adequação do teste do χ^2 para a aferição da significância biológica dos motivos extraídos pelo SMILE.

4.2 Método de Shuffling

Como já foi referido, para efectuar o cálculo da significância estatística, o SMILE recorre à geração de sequências aleatórias com certas restrições baseadas na estatística dos dados originais. Para o efeito é usado um algoritmo de *data shuffling* [7,8].

O método de *shuffling* permite obter um conjunto de dados aleatório com base num conjunto de dados original, garantindo a preservação exacta das frequências de *k-mers* para um dado *k*.

O método de geração pode ser visto como um passeio aleatório sobre um multi-grafo onde os nós são todas as subsequências de tamanho *k-1* presentes na sequência original e cada aresta corresponde a uma transição entre (*k-1)-mers* verificada nos dados. No caso de *k* igual a 1 não é gerado um multi-grafo, sendo feita apenas uma reordenação aleatória dos nucleótidos de cada sequência.

A Figura 4.1 ilustra um exemplo de um multi-grafo gerado a partir da sequência ATTATTTATT, bem como um caminho no grafo que gera a sequência alternativa, TATTATTTAT.

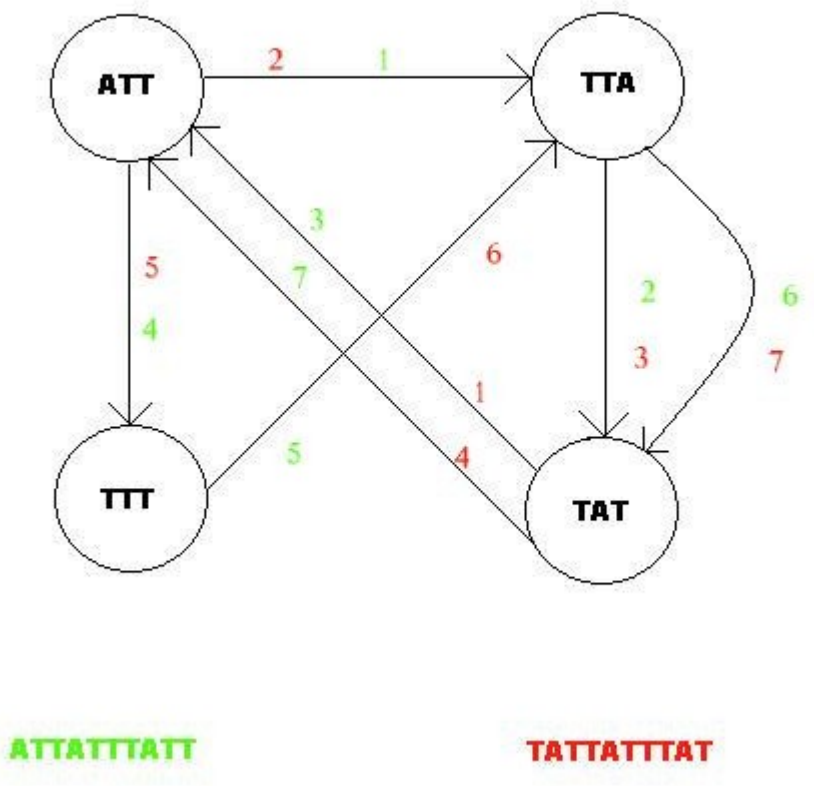


Figura 4.1 – Multi-grafo gerado pela sequência ATTATTTATT para *k*=4, com 2 passeios aleatórios

Para que cada sequência gerada seja válida é necessário que tenha o mesmo tamanho que a sequência original. Sendo assim, é fácil perceber que apenas os caminhos eulerianos garantem a geração de uma sequência com estas características. No SMILE, este processo é repetido para cada sequência, isoladamente.

Para garantir a fiabilidade dos resultados obtidos é efectuado um número elevado de simulações, donde são extraídos os valores médios do quórum ou da abundância de cada motivo a partir dos quais é calculada a significância estatística, no sentido do teste do χ^2.

A Figura 4.2 mostra um histograma com os resultados obtidos em 100 simulações para um motivo extraído do conjunto de dados da *Helicobater pylori*[3] onde foram preservadas as frequências relativas de nucleótidos (*k*=1). Este conjunto de dados é constituído por 308 sequências com comprimentos que variam entre 40 e 300 nucleótidos. Foi efectuada uma procura de motivos simples, com tamanhos não inferiores a 6. Foi admitido 1 erro e foi exigido um quórum de 6%.

O histograma indica a percentagem de simulações em que o motivo em consideração ocorre num determinado número de sequências. Em particular, no exemplo da Figura 4.2, o motivo TTTTAA ocorre num número de sequências entre 282 e 284, em 13% das simulações.

A seta a negro refere-se ao número de ocorrências do motivo observadas no conjunto de dados original. Quanto mais afastada estiver a seta do máximo da curva, maior será a significância estatística. Neste caso, foi atribuída uma significância de 5,93 ao motivo TTTTAA.

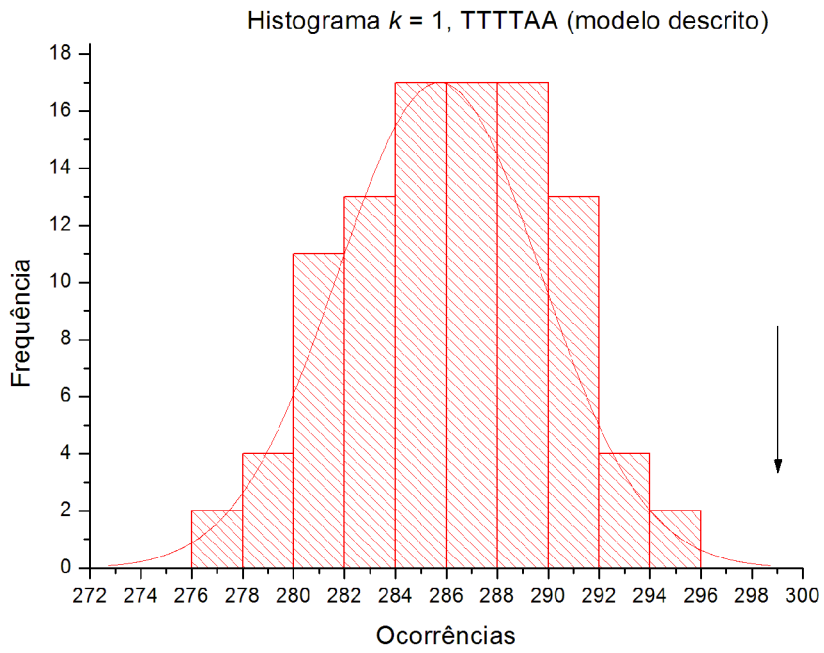


Figura 4.2 – Histograma das ocorrências observadas do motivo TTTTAA em 100 simulações de *shuffling* do conjunto de dados da *H. pylori*

[3] Organismo procariota.

O método de *data shuffling* pode ser pesado do ponto de vista computacional e representa uma grande restrição no desempenho da avaliação de motivos. Em particular, a necessidade de pesquisar todos os motivos extraídos em cada conjunto de dados aleatório gerado constitui uma sério entrave ao bom desempenho deste método.

Devido aos referidos problemas de desempenho opta-se, geralmente, por fazer um número de simulações com geração de sequências aleatórias inferior ao estatisticamente aconselhado, o que pode ter impacto nos resultados.

Outra questão pertinente coloca-se ainda quanto à preservação dos *k-mers*. Uma vez que o *shuffling* é feito sobre cada sequência do conjunto de dados, valores elevados de *k* impõem restrições severas à variabilidade dos conjuntos de dados aleatórios gerados. Assim, para estes valores, o teste estatístico é inútil, uma vez que os conjuntos de dados gerados são praticamente iguais ao conjunto de dados original. Em particular, quando o tamanho dos *k-mers* a preservar é igual ou superior ao tamanho de um motivo, as suas ocorrências são estritamente conservadas no conjunto de dados aleatório, levando a que o valor da significância seja 0. A Figura 4.3 ilustra o facto de a significância estatística tender inevitavelmente para zero à medida que o valor de *k* aumenta. Os valores do gráfico foram igualmente obtidos a partir do conjunto de dados da *H. pylori* descrito anteriormente.

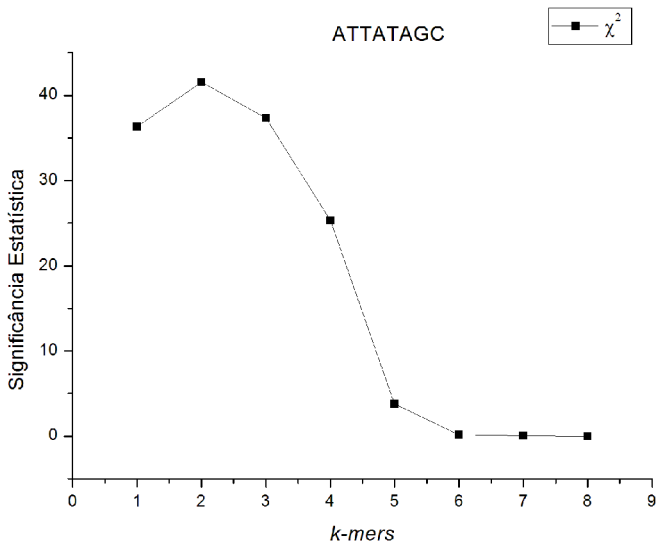


Figura 4.3 – Gráfico da evolução da significância estatística com o aumento do valor de *k* para um motivo de tamanho 8 extraído do conjunto de dados da *H. pylori*

4.3 Método Analítico

Nesta Secção propõe-se um método analítico que tem por objectivo resolver alguns dos problemas do método de *shuffling* descrito na Secção anterior, designadamente, o fraco desempenho computacional e as excessivas restrições devidas à conservação estrita das frequências dos *k-mers*. Assim, pretende-se obter um método computacionalmente eficiente para avaliar a relevância de motivos extraídos pelo SMILE.

Com este método começa-se por abstrair um modelo de *Markov* a partir do conjunto de dados em estudo. Os estados do modelo de *Markov* correspondem a todos os (*k-1)-mers* presentes no conjunto de dados, sendo a probabilidade inicial de cada estado a frequência relativa observada do (*k-1)-mer* correspondente.

A probabilidade de transição entre estados corresponde à razão entre o número de vezes que o (*k-1)-mer* referente ao estado de destino sucede ao (*k-1)-mer* referente ao estado de origem nos dados e o número de vezes que o (*k-1)-mer* correspondente ao estado de origem precede qualquer outro (*k-1)-mer* nas sequências em análise. A Figura 4.5 apresenta um modelo de *Markov* abstraído a partir do conjunto de dados referido na Figura 4.4.

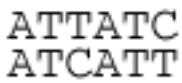
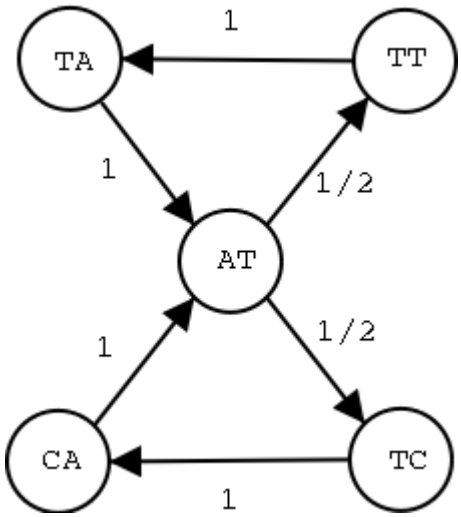


Figura 4.4 – Conjunto de dados exemplo



$$\pi = \begin{pmatrix} 1/5 \\ 1/10 \\ 2/5 \\ 1/5 \\ 1/10 \end{pmatrix} \begin{matrix} TT \\ TA \\ AT \\ TC \\ CA \end{matrix}$$

Figura 4.5 – Exemplo de uma cadeia de *Markov* para *k*=3

Cada motivo em análise é processado de forma a calcular a probabilidade do motivo ser gerado pelo modelo de *Markov* abstraído.

Esta probabilidade é calculada como o produto da probabilidade do estado inicial pela probabilidade de todas as transições necessárias para compor o motivo. Em (4.4), (4.5) e (4.6) apresenta-se o calculo desta probabilidade para o motivo ATTATT segundo a cadeia de *Markov* da Figura 4.5.

$$p(ATTATT) = p(AT)\,p(AT \rightarrow TT)\,p(TT \rightarrow TA)\,p(TA \rightarrow AT)\,p(AT \rightarrow TT) \tag{4.4}$$

$$p(ATTATT) = \frac{2}{5} \times \frac{1}{2} \times 1 \times 1 \times \frac{1}{2} \tag{4.5}$$

$$p(ATTATT) = \frac{1}{10} \tag{4.6}$$

No caso em que *k* é igual 1, não é construída uma cadeia de *Markov*, sendo a probabilidade de ocorrência do motivo o produto das frequências relativas no conjunto de dados dos nucleótidos que o compõem.

Caso sejam admitidos *e* erros basta considerar a soma das probabilidades de todos motivos a uma distância de *Hamming* não superior a *e*, ao invés de se considerar apenas a probabilidade do motivo em análise.

A probabilidade assim calculada será usada para determinar as entradas referentes a dados aleatórios na tabela de contingência descrita na Secção 4.1, para o teste relativo à abundância. A determinação das entradas na tabela quando consideramos o teste relativo ao quórum de um motivo é consideravelmente mais complexa, pelo que não a abordaremos neste texto.

Assim, as entradas P_{random} e A_{random} são calculadas da forma indicada em (4.7) e (4.8), respectivamente.

$$P_{random} = p_{Markov}(m,k) \times (nseqs \times lenseqs - (lenmotif(m) - 1) \times nseqs) \quad (4.7)$$

$$A_{random} = nseqs \times lenseqs - P_{random} \quad (4.8)$$

Nas expressões (4.7) e (4.8),

- $p_{Markov}(m,k)$ refere-se à probabilidade calculada para o motivo *m* a partir de uma cadeia de *Markov* extraída do conjunto de dados para um dado *k*;
- *nseqs* designa o número de sequências do conjunto de dados ;
- *lseqs* designa o comprimento médio das sequências do conjunto de dados;
- *lenmotif(m)* designa o tamanho do motivo *m*.

À semelhança do SMILE, no cálculo de A_{random}, consideramos como número de oportunidades de ocorrência o tamanho do conjunto de dados.

Este método apresenta um desempenho computacional significativamente melhor que o método baseado em *shuffling* uma vez que não necessita de efectuar várias simulações, tendo apenas de processar o conjunto de dados por uma única vez. O cálculo da significância estatística de cada motivo é linear em relação ao seu comprimento.

Por outro lado, o método analítico não enferma das mesmas limitações do que o teste estatístico baseado em simulações, permitindo o cálculo de significância estatística para

valores elevados de *k*. Ao considerar a estatística de todo o conjunto de dados ao invés de cada sequência, isoladamente, este método pode também, à partida, modelar melhor os dados.

Resta acrescentar que o método descrito nesta Secção só permite obter significâncias estatísticas para motivos simples. A realização de cálculos para motivos estruturados é consideravelmente mais complexa e não foi abordada neste trabalho.

5 Geração de código IUPAC

A realização eficiente de pesquisas de consensos em bases de dados biológicas requer uma representação compacta dos dados armazenados. Neste contexto, a possibilidade de representar consensos em código IUPAC assume uma importância fundamental.

O código IUPAC é uma extensão sobre a representação de sequências de DNA. Uma sequência de DNA é constituída por nucleótidos. A codificação IUPAC introduz símbolos que representam um conjunto de possíveis nucleótidos para uma dada posição, permitindo uma representação de sequências degeneradas. Por exemplo, uma purina (A ou G) é representada pela letra R. A descrição completa dos códigos IUPAC encontra-se na Tabela 5.1.

Código	Descrição
A	Adenina
C	Citosina
G	Guanina
T	Timina
U	Uracilo
R	Purina (A ou G)
Y	Pirimidina (C, T, ou U)
M	C ou A
K	T, U, ou G
W	T, U, ou A
S	C ou G
B	C, T, U, ou G (não A)
D	A, T, U, ou G (não C)
H	A, T, U, ou C (não G)
V	A, C, ou G (não T, não U)
N	Qualquer base (A, C, G, T, ou U)

Tabela 5.1 – Código IUPAC

A conversão de um conjunto de sequências de DNA para um conjunto de sequências IUPAC não é uma tarefa trivial. Para um dado conjunto de sequências existem diversas codificações IUPAC possíveis.

O problema da geração de código IUPAC consiste em encontrar o conjunto mínimo de sequências IUPAC que representem um dado conjunto de sequências DNA.

Como o objectivo proposto é a compressão da representação das sequências de nucleótidos, é desejável que a codificação IUPAC seja a mais compacta possível.

A abordagem utilizada foi a de reduzir o problema da compressão IUPAC a um problema conhecido: a minimização de expressões lógicas multi-valor.

Uma expressão lógica multi-valor é uma expressão lógica onde figuram variáveis multi-valor, ou seja, variáveis que necessitam de mais do que um bit para serem representadas.

Para representar um símbolo IUPAC é utilizada uma variável multi-valor de 4 bits. Os símbolos e as respectivas representações estão na Tabela 5.2.

Símbolo IUPAC	Codificação binária
A	1000
T	0100
C	0010
G	0001
R	1001
Y	0110
S	0011
W	1100
K	0101
M	1010
B	0111
D	1101
H	1110
V	1011
N	1111
Espaço	0000

Tabela 5.2 – Codificação binária dos símbolos IUPAC

Uma sequência de nucleótidos ou IUPAC é construída através de um produto de termos, sendo, cada termo, uma variável multi-valor que representa um símbolo da sequência. Por exemplo, a sequência WTA é representada pela expressão lógica (1100)·(0100)·(1000).

Agora que foi estabelecida uma representação para sequências IUPAC através de expressões lógicas multi-valor, resta perceber como, da minimização destas expressões, pode resultar o desejado código IUPAC.

Em (5.1) temos um exemplo simples de minimização lógica.

$$X \cdot Y \cdot Z + X \cdot \overline{Y} \cdot Z = X \cdot Z \qquad (5.1)$$

É fácil verificar que $X \cdot Z$ é uma expressão equivalente. O que se pretende para a compressão IUPAC é algo semelhante.

Por exemplo, a partir das sequências ATA e TTA pretende-se a expressão IUPAC WTA. A simplificação das expressões lógicas é exemplificada sucessivamente em (5.2), (5.3) e (5.4).

A T A + T T A (5.2)

(1000)·(0100)·(1000) + (0100)·(0100)·(1000)

(A ou T) T A (5.3)

((1000)+(0100)) · (0100)·(1000)

W T A (5.4)

(1100) ·(0100) ·(1000)

É fácil perceber, atendendo à escolha criteriosa da codificação, que da minimização de expressões lógicas que representam sequências de nucleótidos, resultam expressões lógicas que representam uma codificação IUPAC das sequências originais.

O problema inicial encontra-se agora reduzido a um problema de minimização de expressões lógicas. Este problema já está resolvido na área da síntese lógica. Um dos programas mais conhecidos, desenvolvido para a resolução deste problema, é o programa *Espresso* [9]. Apesar de ser utilizado, neste contexto, como uma caixa preta é importante perceber o seu funcionamento.

No *Espresso*, as expressões lógicas são representadas num espaço booleano multi-dimensional. Um exemplo é um espaço tridimensional onde podem ser representadas sequências de tamanho 3, tal como está ilustrado na Figura 5.1.

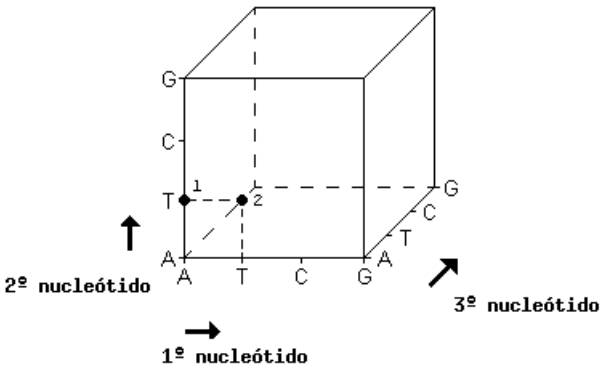


Figura 5.1 – Representação das sequências ATA (1) e TTA (2)

Neste contexto, um cubo define-se como a representação espacial de um produto de termos. Ou seja, para o problema em questão, cada sequência de entrada é um cubo e uma cobertura corresponde a um conjunto de cubos que engloba todos os vértices representados no espaço. Pode observar-se que a dimensão do espaço encontra-se directamente relacionada com o tamanho da sequência que se pretende representar. Para efectuar a minimização, o *Espresso* procura uma cobertura mínima que dará origem ao conjunto de sequências IUPAC de saída.

No caso do espaço da Figura 5.1, existe uma cobertura mínima constituída por um único cubo. A Figura 5.2 ilustra a referida cobertura.

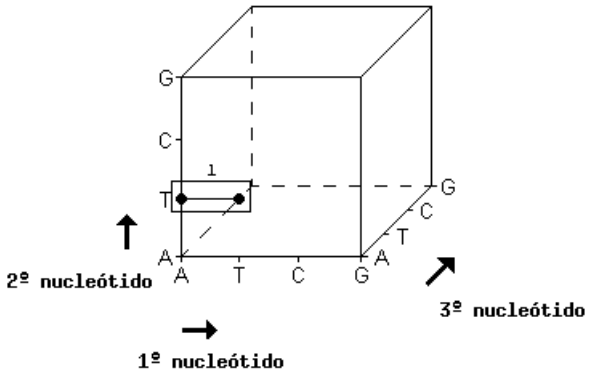


Figura 5.2 – Representação da cobertura WTA (1)

Da cobertura resulta a codificação IUPAC das sequências iniciais.

O processo para descobrir a cobertura mínima é complexo, sendo composto por um conjunto intrincado de passos. Inicialmente, o conjunto de cubos originais é expandido de forma a englobar o maior número de vértices adjacentes. De seguida, os cubos redundantes são removidos e os restantes são reduzidos para que cubram apenas os cubos iniciais. Um cubo é considerado redundante se não cobrir nenhum vértice que não esteja já coberto por outro cubo. Deste processo resulta uma redução contínua do número de cubos necessários para a cobertura. O processo é reiterado até não se verificarem melhorias substanciais na dimensão da cobertura obtida.

Este processo é guiado por heurísticas e é possível demonstrar que produz resultados próximos da solução óptima. O processo está detalhadamente descrito em [9,10].

De modo a automatizar o processo de codificação das sequências de nucleótidos e descodificação das sequências IUPAC foi desenvolvida uma interface para o *Espresso*. Assim, as sequências de entrada são transformadas em expressões lógicas (formato binário), seguindo a correspondência referida na Tabela 5.2. Do mesmo modo, a representação da cobertura mínima obtida pelo *Espresso*, é descodificada e traduzida para sequências IUPAC.

O funcionamento da interface foi testado com recurso a um gerador aleatório de sequências de nucleótidos e a um verificador. O gerador produz um conjunto de sequências, obtidas a partir de uma sequência inicial de nucleótidos. O conjunto de sequências gerado corresponde a várias substituições aleatórias na sequência inicial e constitui o conjunto de entrada que vai ser lido pela ferramenta *Espresso*. O verificador recebe o conjunto de sequências de entrada e o resultado da operação de minimização do *Espresso*, sob a forma de sequências de códigos IUPAC. Deste modo, ao expandir as sequências IUPAC obtidas pelo *Espresso*, através de uma operação trivial de substituições sucessivas, é possível testar a completude e adequação da representação. Assim, diz-se que a representação é completa se a expansão incluir todas as sequências de entrada; por outro lado, diz-se que a representação é adequada se não incluir nenhuma sequência que não esteja no conjunto inicial.

Depois de verificada a completude e adequação das representações obtidas foram feitos testes com dados reais de modo a poder aferir o grau de compressão das sequências introduzidas. Assim, o *Espresso* foi executado tendo como conjunto de entrada sequências, devidamente codificadas, obtidas através de uma execução do SMILE.

As taxas de compressão obtidas variam entre 27.8 e 71.5%.

O gerador de código IUPAC encontra-se neste momento integrado na base de dados ***Dbyeast*** [11], disponibilizando publicamente o serviço de codificação. A Figura 5.3 é uma imagem da interface que permite que os utilizadores da ***Dbyeast*** obtenham código IUPAC a partir de sequências de DNA.

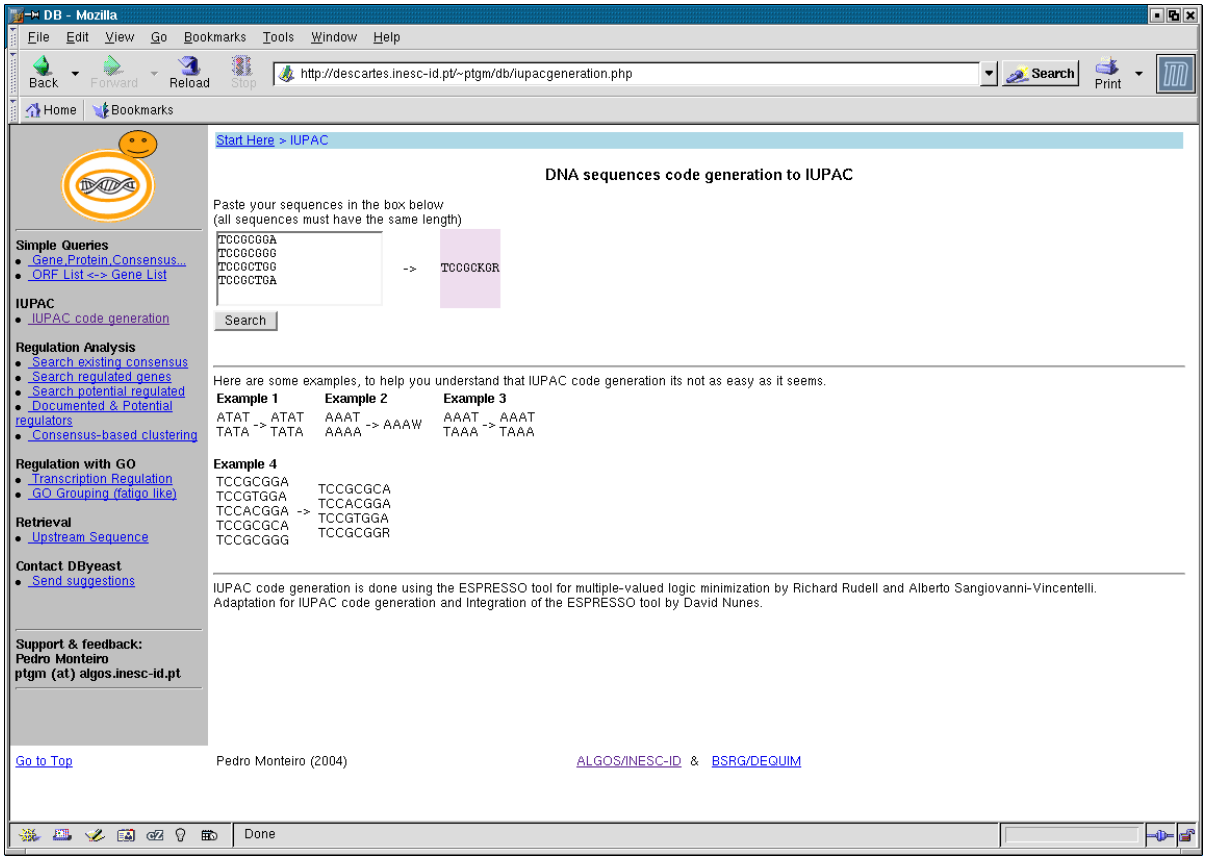


Figura 5.3 – Interface para geração de código IUPAC integrada na ***Dbyeast***

6 Resultados e Conclusões

Neste Capítulo iremos analisar os resultados obtidos tanto para dados sintéticos como para dados reais. Para tal, iremos comparar os valores de significância estatística obtidos para os dois tipos de dados usando tanto o método de *shuffling* como o método analítico. Para efectuar esta comparação é útil referirmo-nos à posição obtida pelos motivos em análise numa lista ordenada por valores decrescentes de significância estatística. Assim, sempre que mencionarmos o valor da posição de um motivo estaremos a referir-nos à posição numa lista assim definida.

6.1 Resultados Obtidos com Dados Sintéticos

Como já foi referido, a utilização de dados sintéticos permite um maior controlo sobre os motivos que estão efectivamente presentes no conjunto de dados. Assim, iremos referir-nos aos motivos que foram introduzidos no conjunto de dados gerado por *motivos especificados* em oposição aos outros motivos igualmente extraídos pelo SMILE que designaremos por *motivos não-especificados*. Nesta Secção iremos considerar o teste de significância estatística relativo à abundância de forma a podermos comparar os resultados obtidos com o método *shuffling* e com o método analítico.

6.1.1 Resultados do teste do χ^2 com o método *shuffling*

Ao especificar um conjunto de dados com os mesmos motivos que estão presentemente descritos para a *H. pylori*, obtivemos, para motivos simples, os resultados apresentados nas Figuras 6.1 e 6.2. O conjunto de dados sintético contém 308 sequências de comprimento 167. Foi efectuada uma procura de motivos simples, com tamanhos compreendidos entre 6 e 8 nucleótidos. Não foram admitidos erros e foi exigido um quórum de 6%. Os motivos especificados encontram-se representados a ponteado. Os restantes motivos correspondem a motivos não-especificados.

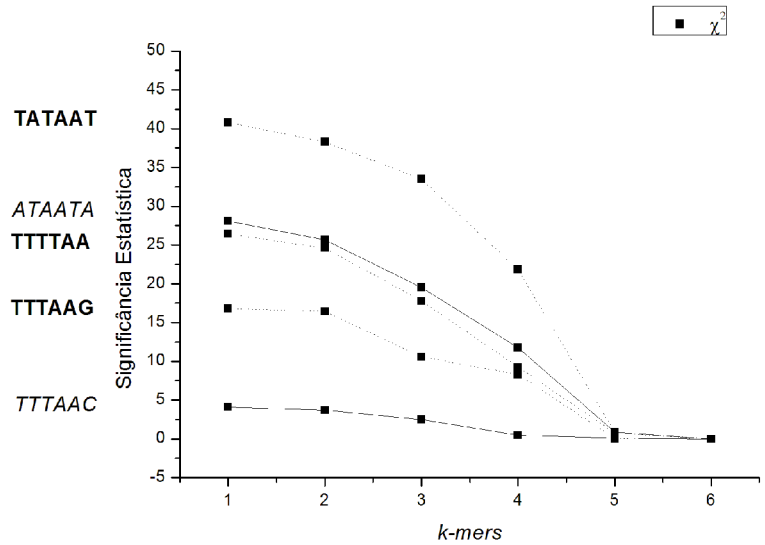


Figura 6.1 – Significância estatística para motivos simples de tamanho 6

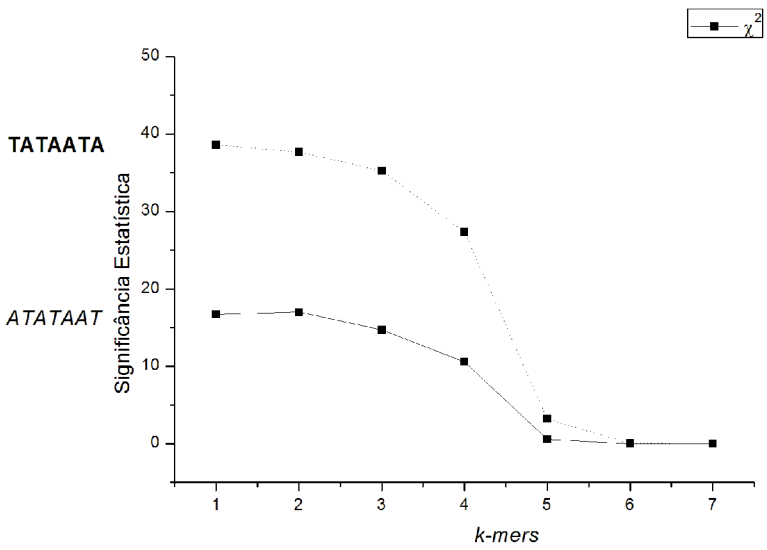


Figura 6.2 – Significância estatística para motivos simples de tamanho 7

Os gráficos não são, no entanto, suficientes para nos dar a percepção da correcta classificação dos motivos, uma vez que não é possível fazer representar no gráfico os 172 motivos extraídos. A Tabela 6.1 indica as posições ocupadas pelos motivos especificados em relação à totalidade dos modelos, para vários valores de *k*.

	TTTAAG	TTTTAA	TATAAT	TATAATA
k=1	5	4	1	2
k=2	6	4	1	2
k=3	6	4	2	1
k=4	6	5	2	1
k=5	78	7	5	1

Tabela 6.1 – Posições dos motivos especificados

Em geral, os motivos especificados aparecem melhor classificados para um baixo valor de *k*. À medida que se consideram valores de *k* crescentes, a significância estatística de todos os motivos aproxima-se de zero, pelas razões já apontadas na Secção 4.2.

É comum observar-se a presença de motivos não-especificados com níveis de significância elevados, especialmente se atendermos às posições da Tabela 6.1.

Contudo, estes bons resultados não se mantêm se forem admitidos erros na procura. Neste caso, embora os motivos mantenham, em geral, elevadas significâncias, entram em competição com os motivos a distância de *Hamming* não superior ao número de erros admitido. Assim, é comum observar-se uma descida significativa na tabela de posições.

Para motivos estruturados os resultados são favoráveis, mesmo admitindo erros, como é ilustrado pelo gráfico da Figura 6.3 em conjugação com os valores de posição apresentados na Tabela 6.2.

Neste exemplo efectuou-se uma procura por motivos estruturados compostos por duas partes com tamanhos entre 6 e 7, com uma distância entre si variando entre 21 e 23 nucleótidos. Admitiu-se 1 erro em cada parte do motivo, mas apenas 1 erro na totalidade do motivo estruturado e exigiu-se um quórum de 6%. Foram extraídos 53 motivos. O motivo especificado encontra-se a ponteado.

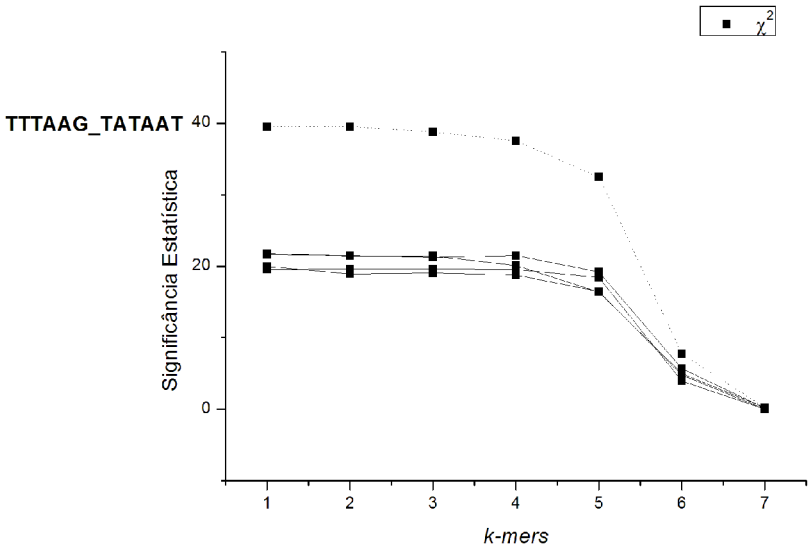


Figura 6.3 – Significância estatística para modelos estruturados em dados sintéticos

	TTTAAG_TATAAT
k=1	4
k=2	4
k=3	17
k=4	63
k=5	162

Tabela 6.2 – Posições do motivo estruturado especificado

Neste exemplo foi necessário admitir a existência de erros para que o motivo fosse encontrado pelo SMILE. Não existem motivos sem erros que cumpram as condições de procura.

Esta boa prestação dos motivos estruturados está relacionada com o facto de o *shuffling* raramente preservar as distâncias relativas de cada uma das partes que constituem o motivo, fazendo com que estes ocorram poucas vezes nas simulações. Assim, neste exemplo, o motivo

estruturado especificado aparece claramente destacado. Estes resultados parecem sugerir que a utilização deste teste estatístico se adequa à avaliação de modelos estruturados. No entanto, nem todos os promotores exibem esta estrutura, sendo necessário um método mais geral para fazer esta avaliação.

Como veremos em seguida, os resultados aparentemente animadores, mesmo para motivos simples, não se verificam quando analisamos dados reais. Esta evidência está relacionada com o facto de o gerador de dados sintéticos gerar dados aleatórios com equiprobabilidade de nucleótidos. Nos dados reais isto não se verifica, existindo, normalmente, uma clara abundância de A's e T's na região promotora. Do mesmo modo, cada conjunto de dados exibe preferência por certos dímeros ou trímeros, que ocorrem com muito mais frequência que outros.

Esta limitação podia ser mitigada pela incorporação destas correlações na geração de dados sintéticos. Mas embora seja relativamente simples estabelecer uma preferência por certos nucleótidos, a geração não-equiprovável de *k-mers* genéricos constitui, só por si, um problema complexo pelo que não foi abordado neste trabalho.

6.1.2 Resultados do teste do χ^2 com o método analítico

Os dados sintéticos foram igualmente analisados recorrendo ao método analítico que, como já foi referido, tem a vantagem de ser mais eficiente do ponto de vista computacional e de poder explorar valores mais elevados de *k*. As Figura 6.4 e 6.5 apresentam a evolução da significância estatística dos mesmos motivos considerados nas Figuras 6.2 e 6.3, mas recorrendo ao método analítico.

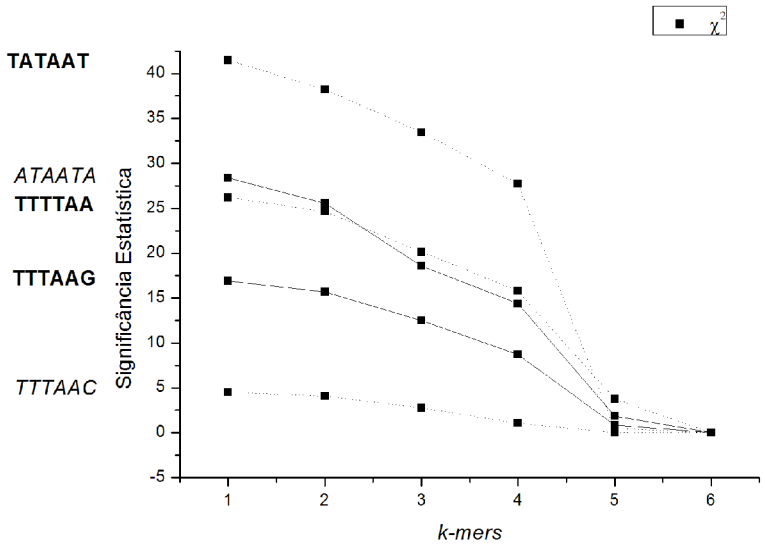


Figura 6.4 – Significância estatística de motivos simples de tamanho 6 usando o método analítico

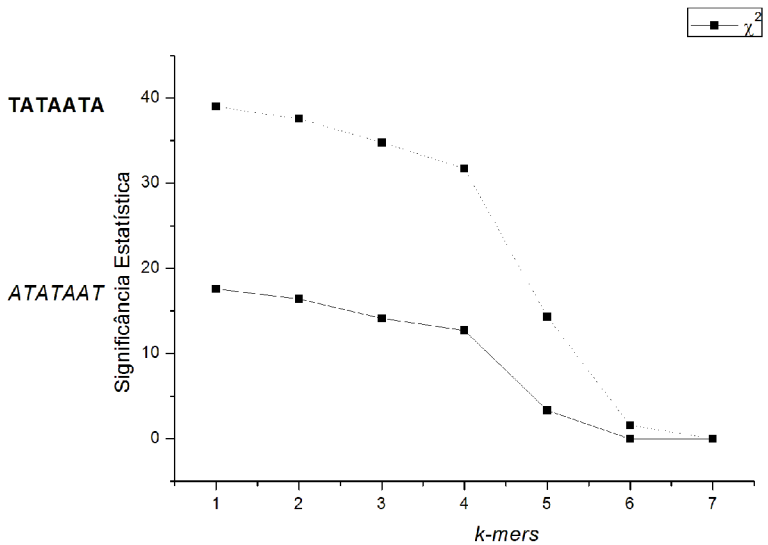


Figura 6.5 – Significância estatística de motivos simples de tamanho 7 usando o método analítico

A Tabela 6.3 indica as posições dos motivos especificados para a significância estatística calculada por via do método analítico.

	TTTAAG	TTTTAA	TATAAT	TATAATA
k=1	6	4	1	2
k=2	6	4	1	2
k=3	6	3	2	1
k=4	6	3	2	1
k=5	55	3	2	1

Tabela 6.3 – Posições dos modelos especificados

São óbvias as semelhanças entre os resultados apresentados na Tabela 6.3 e os presentes na Tabela 6.1. Isto confirma a possibilidade de usar o método analítico como alternativa ao método de *shuffling*.

Sabemos igualmente que, no âmbito do método analítico, quando *k* atinge um valor igual ao comprimento do motivo em análise, a significância estatística anula-se. Isto tem a ver com o facto de a probabilidade de ocorrência de um *k-mer* ser muito próxima da sua frequência relativa no conjunto de dados. Assim sendo, os valores das entradas da tabela de contingência relativas aos dados aleatórios serão inevitavelmente iguais às dos dados reais, resultando deste facto uma significância estatística nula.

Por outro lado, à medida que o *k* aumenta, os *k-mers* introduzidos pelos motivos especificados no conjunto de dados que, de resto, é aleatório vão fazer com que a cadeia de *Markov* gerada faça convergir as probabilidades de ocorrência de cada motivo com a sua frequência relativa efectiva no conjunto de dados. Assim, assiste-se ao mesmo fenómeno de convergência para zero da significância estatística que assinalámos para o método *shuffling*.

Pela análise dos resultados, podemos ainda concluir que, em geral, um valor de *k* demasiado pequeno não modela suficientemente bem os dados, daqui resulta que alguns motivos não especificados obtenham melhores classificações que motivos especificados. Em suma, deveremos preferir valores de *k* intermédios.

6.2 Resultados Obtidos com Dados Reais

Com este teste pretendemos verificar se uma sequência classificada como estatisticamente relevante também pode ser considerada biologicamente relevante.

Para fazer o estudo da análise estatística em dados reais, foram recolhidos 16 conjuntos de dados com regiões promotoras correspondentes a genes da levedura *Saccharomyces cerevisiae*, que é um organismo eucariota. Cada conjunto contém aproximadamente 10 sequências. As sequências contêm consensos obtidos por via experimental e descritos na literatura. A descrição completa de cada conjunto de dados pode ser vista no Anexo A.

Para cada conjunto de dados foi feita uma procura de motivos de tamanhos variando entre 4 e 11 nucleótidos, para valores de quórum de 60%, 75%, 80% e 100%. Foram permitidos 0 e 1 erros, em ensaios sucessivos e fez-se ainda variar o valor de *k* entre 1 e 4.

Foi posteriormente analisada a significância estatística, calculada pelo SMILE, no que diz respeito ao quórum.

Os consensos descritos obtiveram boas classificações para algumas das condições testadas. Em particular, quando o quórum exigido se encontra próximo da frequência efectiva dos motivos nos dados e quando não são admitidos erros. Os resultados são também melhores para motivos de tamanho superior a 7. Em suma, obtêm-se bons resultados quando as condições favorecem os motivos com as características dos consensos descritos.

Estes resultados não são muito animadores porque para os atingir é necessário um conhecimento prévio da frequência e tamanho dos motivos. Ao considerar a generalidade dos ensaios efectuados, verifica-se que as classificações são muito baixas, com baixas significâncias estatísticas e com valores de posição frequentemente abaixo dos primeiros 1 000 motivos. No Anexo A, apresentamos os resultados obtidos com estes ensaios.

A par desta análise, foi considerada a significância estatística obtida a respeito da abundância. Deste modo, é possível estabelecer uma comparação directa entre o método de *shuffling* e o método analítico para o cálculo desta métrica.

De seguida, apresentamos os resultados desta análise.

6.2.1 Resultados do teste do χ^2 com o método *shuffling*

De modo geral, os resultados para o teste estatístico referente à abundância representam uma melhoria em relação ao teste referente ao quórum.

Começamos por apresentar os gráficos referentes a um conjunto de dados cujos motivos permanecem mal classificados, isto é, o motivo é biologicamente relevante, mas não é considerado estatisticamente relevante. O conjunto de dados em causa, o MSN4, tem um motivo muito abundante (CCCCT), condicionando a estatística dos dados. Isto, conjugado com o seu pequeno tamanho (5 nucleótidos) resulta na inevitável baixa significância estatística, uma vez que aparece sistematicamente na geração aleatória de sequências feita em cada simulação. As Figuras 6.6 e 6.7 ilustram a evolução da significância estatística e da posição do referido motivo considerando erro 0 e 1, respectivamente, para um quórum de 60%.

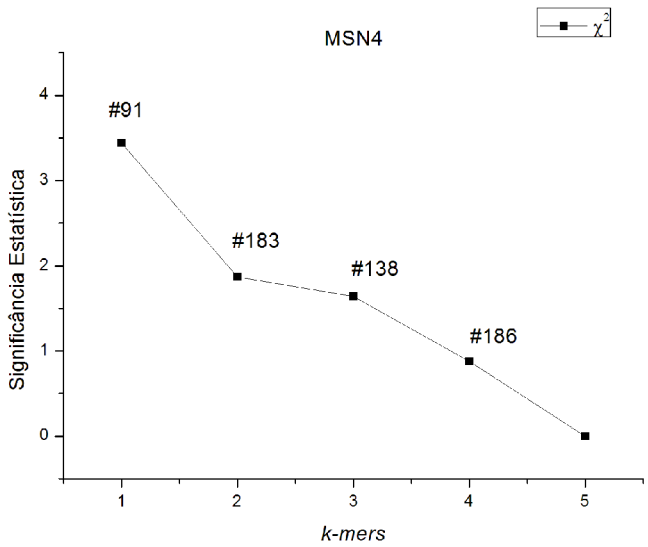


Figura 6.6 – Evolução da significância estatística relativa à abundância com o valor de *k* para o conjunto de dados MSN4 com 0 erros

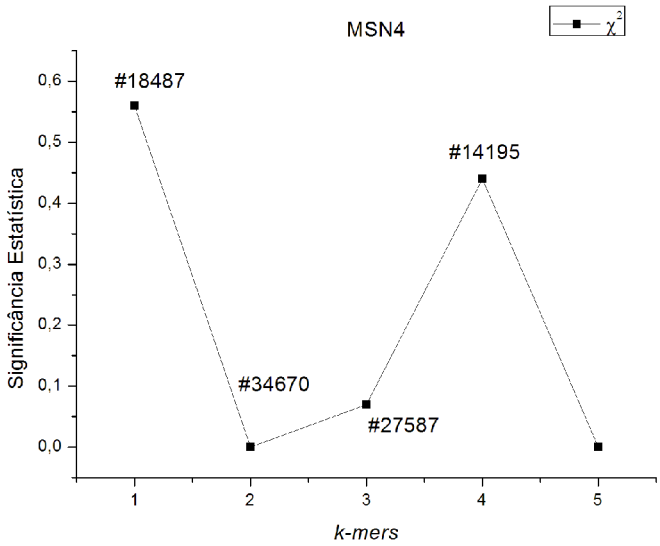


Figura 6.7 – Evolução da significância estatística relativa à abundância com o valor de *k* para o conjunto de dados MSN4 com 1 erro

Os resultados da classificação são claramente maus para um consenso descrito. A estranha subida para *k* igual a 4 está relacionada com uma pequena flutuação sem significado, uma vez que estamos a considerar valores de significância estatística muito baixos. A aparente melhor classificação quando consideramos erro 0 está relacionada com o facto de que quando não se admitem erros são consequentemente extraídos menos motivos. Em particular, com erro 0 são extraídos 1 316 motivos enquanto que com erro 1 são extraídos 35 258 motivos.

De seguida, apresentamos os resultados relativos a um outro conjunto de dados (STE12) que evidencia resultados mais favoráveis para erro 0. As Figuras 6.8 e 6.9 apresentam a evolução da significância estatística do motivo descrito (TGAAACA), para 0 e 1 erros, respectivamente, com um quórum de 100%.

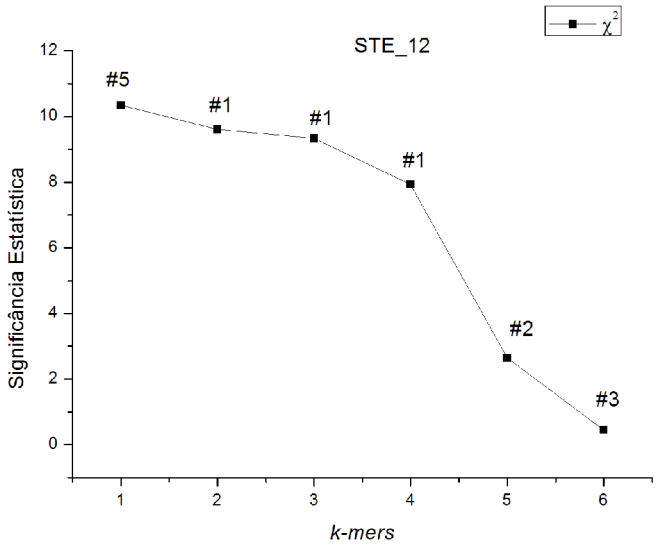


Figura 6.8 – Evolução da significância estatística relativa à abundância com o valor de *k* para o conjunto de dados STE12 com 0 erros

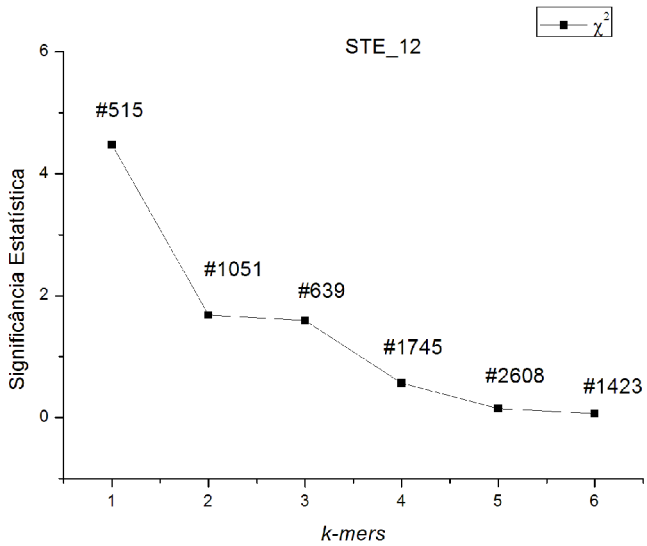


Figura 6.9 – Evolução da significância estatística relativa à abundância com o valor de *k* para o conjunto de dados STE12 com 1 erro

Resta referir que, para este conjunto de dados, foram extraídos 267 motivos com erro 0 e 7 705 motivos para erro 1.

O facto de os motivos serem abundantes em relação ao número de sequências consideradas faz com que o teste relativo à abundância tenha um poder de resolução consideravelmente maior do que o teste relativo ao quórum. Por esta razão deveremos sempre preferir o primeiro em casos semelhantes.

Por outro lado, os resultados obtidos continuam a pôr em causa a generalidade do método uma vez que é difícil identificar as situações em que este produz bons resultados sobretudo quando somos confrontados com um desconhecimento *a priori* do conjunto de dados.

6.2.2 Resultados do teste do χ^2 com o método analítico

Resta-nos comparar os resultados obtidos na subsecção anterior com os obtidos pela via analítica. Nos gráficos apresentados nesta subsecção, que representam a evolução da significância estatística com o crescimento do tamanho dos *k-mers* considerados, estão também assinaladas as posições ocupadas pelos motivos.

As Figuras 6.10 e 6.11 apresentam os resultados referentes ao MSN4.

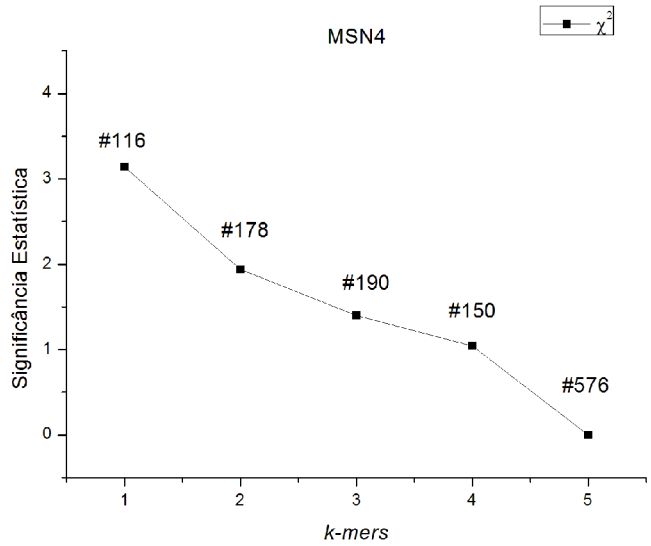


Figura 6.10 – Evolução da significância estatística relativa à abundância com o valor de *k* para o conjunto de dados MSN4 com 0 erros

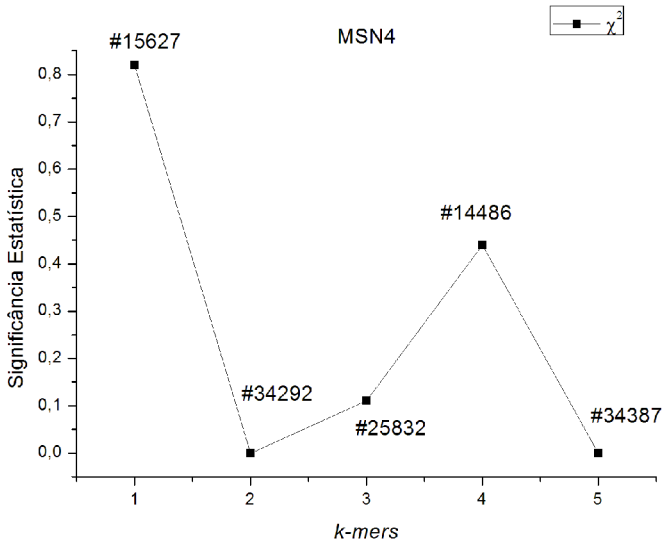


Figura 6.11 – Evolução da significância estatística relativa à abundância com o valor de *k* para o conjunto de dados MSN4 com 1 erro

Os resultados para o MSN4 não sofrem alterações assinaláveis em relação ao método anterior. Novamente, as características do conjunto de dados limitam a possibilidade de evidenciar este

motivo por via deste teste estatístico o que é tanto mais grave quanto o facto de este ser um dos motivos cuja significância biológica se encontra melhor estabelecida.

As Figuras 6.12 e 6.13 apresentam os resultados referentes ao conjunto de dados STE12.

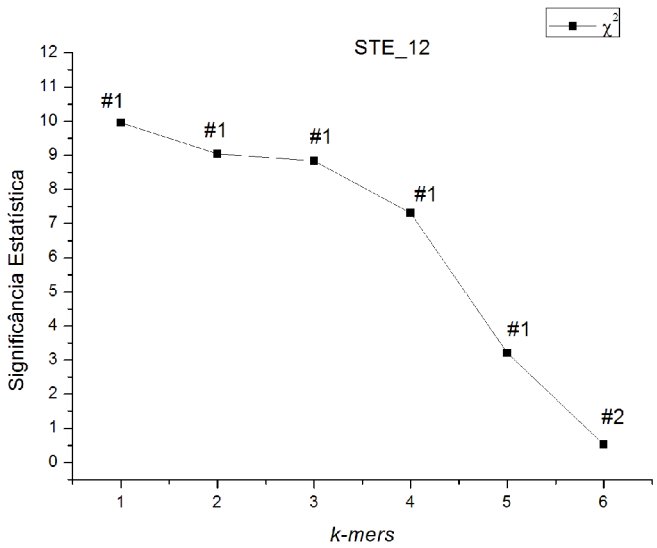


Figura 6.12 – Evolução da significância estatística relativa à abundância com o valor de *k* para o conjunto de dados STE12 com 0 erros

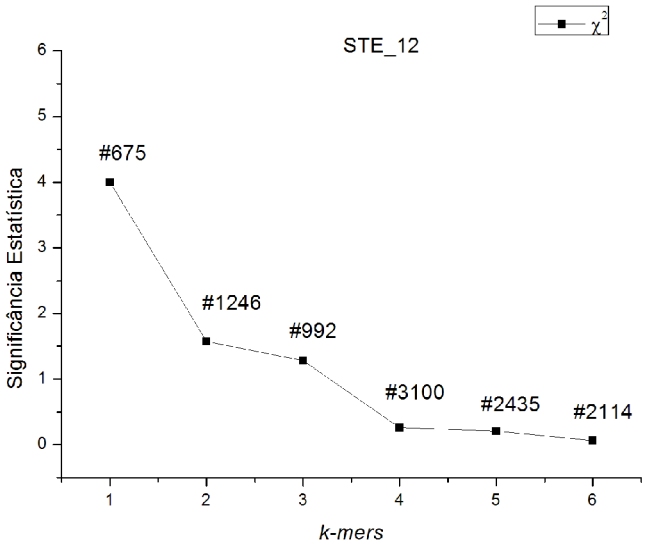


Figura 6.13 – Evolução da significância estatística relativa à abundância com o valor de *k* para o conjunto de dados STE12 com 1 erro

Os dados apresentados não evidenciam melhorias assinaláveis em relação aos resultados obtidos por via de *shuffling*, excepto a ocorrência de melhorias ocasionais da classificação quando não se admitem erros. Estas diferenças podem ser atribuídas ao facto de o método analítico não conservar estritamente os *k-mers* e de analisar o conjunto de dados na sua totalidade, por oposição ao *shuffling* que considera cada sequência isoladamente.

No Anexo B apresentamos a totalidade dos resultados desta análise comparativa. No Anexo C é apresentada uma listagem dos resultados do cálculo do valor de significância para valores de *k* até ao tamanho dos consensos descritos.

A grande vantagem deste método é, sem dúvida, o incomparável ganho a nível de desempenho computacional. Para ilustrar esta afirmação, basta dizer que o cálculo da significância estatística para todos os conjuntos de dados considerados demorou, com o método de *shuffling*, mais de 30 horas de processamento. O método analítico realizou todos os cálculos em menos de 30 minutos.

No entanto, os resultados voltam a pôr em causa a adequação do teste estatístico do χ^2 para aferir a significância biológica, pela dificuldade de sintetizar um método genérico para destacar motivos relevantes de entre os extraídos pelo SMILE.

6.3 Conclusões

Ao considerar os resultados apresentados é inevitável concluir que a significância atribuída pelo teste estatístico do χ^2 aos motivos extraídos pelo SMILE não reflecte a sua significância biológica.

A dificuldade em distinguir motivos relevantes da grande quantidade de motivos extraídos põe em causa a utilidade prática do algoritmo, na generalidade dos casos.

Para além desta questão, outras limitações têm sido apontadas por biólogos, designadamente, o facto de o algoritmo não conseguir pesquisar motivos na cadeia de DNA inversa, onde os factores de transcrição podem igualmente ligar-se, e o facto de a forma como o SMILE modela a ocorrência de erros nos motivos não ser compatível com a necessidade de poder indicar posições específicas onde erros são admitidos.

No entanto, o SMILE continua a exibir resultados interessantes para motivos estruturados, o que constitui, sem dúvida, a mais-valia fundamental do algoritmo. Tudo indica que o método analítico pode ser igualmente adaptado para calcular a significância de motivos estruturados.

Parece-nos claro que uma eficaz extracção de motivos relevantes passará, necessariamente, pela conjugação com outros dados, para além da simples sequência de nucleótidos. Em particular, a disponibilidade de informação acerca de co-regulação de genes poderá guiar a extracção de motivos permitindo exigir um quórum mais alto, dada a confiança na existência de factores de transcrição comuns.

Por outro lado, a utilização de modelos probabilísticos em vez de meras colecções de motivos poderá ser uma abordagem mais natural para o problema permitindo, simultaneamente, uma melhor modelação da realidade biológica, ao indicar uma probabilidade de ocorrência de cada nucleótido para uma dada posição no modelo. O número de modelos extraídos seria também francamente reduzido, uma vez que um único modelo probabilístico pode representar, com expressividade acrescida, um grande número de motivos.

7 Referências

[1] A. Vanet, L. Marsan e M. F. Sagot, Promoter sequence and algorithmical methods for identifying them, *Research in Microbiology*, 150, pp. 779-799, 1999

[2] J.D. Watson, F.H. Crick, Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid, *Nature*, Number 4356, 1953

[3] T. Werner, Models for Prediction and Recognition of Eukaryotic Promoters, *Mammalian Genome*, Vol. 10, pp. 168-175, 1999

[4] L. Marsan e M. F. Sagot, Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identication, *Journal of Computational Biology*, 7, pp. 345-362, 2000.

[5] L. Marsan, *Inférence de motifs structurés: algorithmes et outils appliqués à la détection de sites de fixation dans des séquences génomiques*, PhD Thesis, Université de Marne-la-Vallée, 2002.

[6] B. Flannery, S. Teukolsky, W. Press e W. Vetterling, *Numerical Recipes in C++* , Cambridge University Press, 2002

[7] S. Altschul e B. Erickson, Significance of Nucleotide Sequence Alignments: A method of Random Sequence Permutation That Preserves Dinucleotide and Codon Usage, *Molecular Biology Evolution*, 2, pp. 526-538, 1985

[8] D. Kandel, Y. Matias, R. Unger e P. Winkler, Shuffling Biological Sequences, *Discrete Appl. Math.*, 71, pp. 171-185, 1996

[9] R. Rudell e A. Sangiovanni-Vincentelli, Multiple-Valued Minimization for PLA Optimization, *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. CAD-6, Nº 5, pp. 727-751, September 1987

[10] R. Brayton, G. Hachtel, C. McMullen e A. Sangiovanni-Vincentelli, *Logic Minimization Algorithms for VLSI Synthesis*, Kluwer Academic Publishers, 1985.

[11] P. Monteiro, Dbyeast, [http://descartes.inesc-id.pt/~ptgm/db/], 2004.

Anexo A – Resultados do Cálculo da Significância Estatística referente ao Quórum para Conjuntos de Dados de *Saccharomyces cerevisiae*

Foi feita uma procura, usando o SMILE, em vários conjuntos de dados.

A parametrização usada fez variar os seguintes parâmetros:

k = 1,...,4
erro = 0,1
quórum = 60, 75, 80, 100

A procura foi feita para motivos com comprimento de 4 a 10, excepto no conjunto de dados ZAP1 onde a procura foi feita para tamanhos entre 4 e 11.

De seguida apresentamos os resultados para os motivos procurados em relação a cada valor dos parâmetros.

Nota: o campo "posicao" refere-se à posição ocupada pelo motivo na tabela com a estatística das sequências com pelo menos uma ocorrência do motivo, ordenada por chi2.

```
11_2_ZAP1

Motivo: ACCTTGAAGGT     ACCCTAAAGGT     ACCTTAAAGGT
Frequência: 4/5 ~80%
Tamanho das Sequências: 1000

Erro = 0
      Quorum = 100%
            314 motivos encontrados
      Quorum = 80%
            670 motivos encontrados
      Quorum = 75%
            670 motivos encontrados
      Quorum = 60%
            1306 motivos encontrados
Erro = 1
      Quorum = 100%
            8636 motivos encontrados
            k = 1 posicao = 2 chi2 = 9.8  (ACCTTAAAGGT)
            k = 2 posicao = 2 chi2 = 9.76 (ACCTTAAAGGT)
            k = 3 posicao = 2 chi2 = 9.65 (ACCTTAAAGGT)
            k = 4 posicao = 2 chi2 = 9.84 (ACCTTAAAGGT)
      Quorum = 80%
            17722 motivos encontrados
            k = 1 posicao = 2 chi2 = 6.51 (ACCCTAAAGGT)
            k = 1 posicao = 4 chi2 = 9.65 (ACCTTAAAGGT)
            k = 2 posicao = 2 chi2 = 6.59 (ACCCTAAAGGT)
            k = 2 posicao = 8 chi2 = 9.65 (ACCTTAAAGGT)
            k = 3 posicao = 2 chi2 = 6.67 (ACCCTAAAGGT)
            k = 3 posicao = 4 chi2 = 9.69 (ACCTTAAAGGT)
            k = 4 posicao = 2 chi2 = 6.59 (ACCCTAAAGGT)
            k = 4 posicao = 3 chi2 = 9.84 (ACCTTAAAGGT)
      Quorum = 75%
            17722 motivos encontrados
```

```
            k = 1 posicao = 4 chi2 = 6.51 (ACCCTAAAGGT)
            k = 1 posicao = 2 chi2 = 9.88 (ACCTTAAAGGT)
            k = 2 posicao = 3 chi2 = 6.59 (ACCCTAAAGGT)
            k = 2 posicao = 2 chi2 = 9.92 (ACCTTAAAGGT)
            k = 3 posicao = 2 chi2 = 6.55 (ACCCTAAAGGT)
            k = 3 posicao = 3 chi2 = 9.76 (ACCTTAAAGGT)
            k = 4 posicao = 3 chi2 = 6.55 (ACCCTAAAGGT)
            k = 4 posicao = 5 chi2 = 9.84 (ACCTTAAAGGT)
      Quorum = 60%
            34907 motivos encontrados
            k = 1 posicao = 11      chi2 = 6.55 (ACCCTAAAGGT)
            k = 1 posicao = 5       chi2 = 9.96 (ACCTTAAAGGT)
            k = 1 posicao = 13      chi2 = 4.21 (ACCTTGAAGGT)
            k = 2 posicao = 15      chi2 = 6.47 (ACCCTAAAGGT)
            k = 2 posicao = 12      chi2 = 9.69 (ACCTTAAAGGT)
            k = 2 posicao = 7       chi2 = 4.21 (ACCTTGAAGGT)
            k = 3 posicao = 9       chi2 = 6.55 (ACCCTAAAGGT)
            k = 3 posicao = 8       chi2 = 9.84 (ACCTTAAAGGT)
            k = 3 posicao = 27      chi2 = 4.14 (ACCTTGAAGGT)
            k = 4 posicao = 10      chi2 = 6.51 (ACCCTAAAGGT)
            k = 4 posicao = 5       chi2 = 9.84 (ACCTTAAAGGT)
            k = 4 posicao = 40      chi2 = 4.07 (ACCTTGAAGGT)


5_0_msn2_4

Motivo: CCCCT
Frequência: 9/13 ~69,2%
Tamanho das Sequências: 1000

Erro = 0
      Quorum = 100%
            161 motivos encontrados
      Quorum = 80%
            381 motivos encontrados
      Quorum = 75%
            502 motivos encontrados
      Quorum = 60%
            826 motivos encontrados
            k = 1 posicao = 43      chi2 = 3.3  (CCCCT)
            k = 2 posicao = 42      chi2 = 2.02 (CCCCT)
            k = 3 posicao = 48      chi2 = 1.42 (CCCCT)
            k = 4 posicao = 122     chi2 = 0.28 (CCCCT)
Erro = 1
      Quorum = 100%
            4877 motivos encontrados
            k = 1 posicao = 3089    chi2 = 0.02 (CCCCT)
            k = 2 posicao = 2781    chi2 = 0.05 (CCCCT)
            k = 3 posicao = 3148    chi2 = 0.02 (CCCCT)
            k = 4 posicao = 2516    chi2 = 0.07 (CCCCT)
      Quorum = 80%
            10050 motivos encontrados
            k = 1 posicao = 8449    chi2 = 0    (CCCCT)
            k = 2 posicao = 5470    chi2 = 0.09 (CCCCT)
            k = 3 posicao = 5867    chi2 = 0.04 (CCCCT)
            k = 4 posicao = 5131    chi2 = 0.09 (CCCCT)
```

```
      Quorum = 75%
            13080 motivos encontrados
            k = 1 posicao = 7539    chi2 = 0.04 (CCCCT)
            k = 2 posicao = 8129    chi2 = 0.02 (CCCCT)
            k = 3 posicao = 7900    chi2 = 0.03 (CCCCT)
            k = 4 posicao = 7385    chi2 = 0.05 (CCCCT)
      Quorum = 60%
            21544 motivos encontrados
            k = 1 posicao = 12673   chi2 = 0.04 (CCCCT)
            k = 2 posicao = 11871   chi2 = 0.08 (CCCCT)
            k = 3 posicao = 12731   chi2 = 0.04 (CCCCT)
            k = 4 posicao = 11754   chi2 = 0.07 (CCCCT)


5_0_MSN4

Motivo: CCCCT
Frequência: 3/5 ~60%
Tamanho das Sequências: 1000

Erro = 0
      Quorum = 100%
            326 motivos encontrados
      Quorum = 80%
            682 motivos encontrados
            k = 1 posicao = 169     chi2 = 1.09 (CCCCT)
            k = 2 posicao = 238     chi2 = 0.73 (CCCCT)
            k = 3 posicao = 223     chi2 = 0.57 (CCCCT)
            k = 4 posicao = 331     chi2 = 0.1  (CCCCT)
      Quorum = 75%
            682 motivos encontrados
            k = 1 posicao = 136     chi2 = 1.31 (CCCCT)
            k = 2 posicao = 178     chi2 = 0.85 (CCCCT)
            k = 3 posicao = 245     chi2 = 0.42 (CCCCT)
            k = 4 posicao = 331     chi2 = 0.1  (CCCCT)
      Quorum = 60%
            1316 motivos encontrados
            k = 1 posicao = 303     chi2 = 1.7  (CCCCT)
            k = 2 posicao = 505     chi2 = 0.75 (CCCCT)
            k = 3 posicao = 457     chi2 = 0.59 (CCCCT)
            k = 4 posicao = 613     chi2 = 0.14 (CCCCT)
Erro = 1
      Quorum = 100%
            9022 motivos encontrados
            k = 1 posicao = 7955    chi2 = 0    (CCCCT)
            k = 2 posicao = 7901    chi2 = 0    (CCCCT)
            k = 3 posicao = 7846    chi2 = 0    (CCCCT)
            k = 4 posicao = 7711    chi2 = 0    (CCCCT)
      Quorum = 80%
            17951 motivos encontrados
            k = 1 posicao = 16310   chi2 = 0    (CCCCT)
            k = 2 posicao = 16251   chi2 = 0    (CCCCT)
            k = 3 posicao = 16207   chi2 = 0    (CCCCT)
            k = 4 posicao = 16091   chi2 = 0    (CCCCT)
      Quorum = 75%
            17951 motivos encontrados
            k = 1 posicao = 16308   chi2 = 0    (CCCCT)
```

```
            k = 2 posicao = 16242   chi2 = 0    (CCCCT)
            k = 3 posicao = 16212   chi2 = 0    (CCCCT)
            k = 4 posicao = 16063   chi2 = 0    (CCCCT)
      Quorum = 60%
            35258 motivos encontrados
            k = 1 posicao = 32818   chi2 = 0    (CCCCT)
            k = 2 posicao = 32757   chi2 = 0    (CCCCT)
            k = 3 posicao = 32822   chi2 = 0    (CCCCT)
            k = 4 posicao = 32656   chi2 = 0    (CCCCT)


5_1_gcr1

Motivo: CWTCC
Frequência: 10/10 ~100%
Tamanho das Sequências: 1000

Erro = 0
      Quorum = 100%
            220 motivos encontrados
      Quorum = 80%
            517 motivos encontrados
            k = 1 posicao = 52      chi2 = 2.54 (CTTCC)
            k = 2 posicao = 87      chi2 = 1.46 (CTTCC)
            k = 3 posicao = 164     chi2 = 0.67 (CTTCC)
            k = 4 posicao = 236     chi2 = 0.04 (CTTCC)
      Quorum = 75%
            517 motivos encontrados
            k = 1 posicao = 77      chi2 = 2.46 (CTTCC)
            k = 2 posicao = 83      chi2 = 1.52 (CTTCC)
            k = 3 posicao = 99      chi2 = 0.98 (CTTCC)
            k = 4 posicao = 241     chi2 = 0.03 (CTTCC)
      Quorum = 60%
            963 motivos encontrados
            k = 1 posicao = 625     chi2 = 0.03 (CATCC)
            k = 1 posicao = 128     chi2 = 2.84 (CTTCC)
            k = 2 posicao = 644     chi2 = 0.03 (CATCC)
            k = 2 posicao = 170     chi2 = 1.77 (CTTCC)
            k = 3 posicao = 490     chi2 = 0.15 (CATCC)
            k = 3 posicao = 256     chi2 = 1.07 (CTTCC)
            k = 4 posicao = 522     chi2 = 0    (CATCC)
            k = 4 posicao = 497     chi2 = 0.04 (CTTCC)
Erro = 1
      Quorum = 100%
            6121 motivos encontrados
            k = 1 posicao = 5072    chi2 = 0    (CATCC)
            k = 1 posicao = 5313    chi2 = 0    (CTTCC)
            k = 2 posicao = 5049    chi2 = 0    (CATCC)
            k = 2 posicao = 5287    chi2 = 0    (CTTCC)
            k = 3 posicao = 4984    chi2 = 0    (CATCC)
            k = 3 posicao = 5236    chi2 = 0    (CTTCC)
            k = 4 posicao = 4894    chi2 = 0    (CATCC)
            k = 4 posicao = 5167    chi2 = 0    (CTTCC)
      Quorum = 80%
            13644 motivos encontrados
            k = 1 posicao = 11732   chi2 = 0    (CATCC)
            k = 1 posicao = 11981   chi2 = 0    (CTTCC)
```

```
            k = 2 posicao = 11792   chi2 = 0    (CATCC)
            k = 2 posicao = 12044   chi2 = 0    (CTTCC)
            k = 3 posicao = 11736   chi2 = 0    (CATCC)
            k = 3 posicao = 11985   chi2 = 0    (CTTCC)
            k = 4 posicao = 11644   chi2 = 0    (CATCC)
            k = 4 posicao = 11917   chi2 = 0    (CTTCC)
      Quorum = 75%
            13644 motivos encontrados
            k = 1 posicao = 11712   chi2 = 0    (CATCC)
            k = 1 posicao = 11966   chi2 = 0    (CTTCC)
            k = 2 posicao = 11791   chi2 = 0    (CATCC)
            k = 2 posicao = 12027   chi2 = 0    (CTTCC)
            k = 3 posicao = 11729   chi2 = 0    (CATCC)
            k = 3 posicao = 11981   chi2 = 0    (CTTCC)
            k = 4 posicao = 11649   chi2 = 0    (CATCC)
            k = 4 posicao = 11920   chi2 = 0    (CTTCC)
      Quorum = 60%
            25804 motivos encontrados
            k = 1 posicao = 22732   chi2 = 0    (CATCC)
            k = 1 posicao = 22973   chi2 = 0    (CTTCC)
            k = 2 posicao = 22881   chi2 = 0    (CATCC)
            k = 2 posicao = 23121   chi2 = 0    (CTTCC)
            k = 3 posicao = 22916   chi2 = 0    (CATCC)
            k = 3 posicao = 23164   chi2 = 0    (CTTCC)
            k = 4 posicao = 22864   chi2 = 0    (CATCC)
            k = 4 posicao = 23141   chi2 = 0    (CTTCC)


6_0_MIG1

Motivo: GCGGGG
Frequência: 1/6 ~16,7%
Tamanho das Sequências: 1000

Erro = 0
      Quorum = 100%
            290 motivos encontrados
      Quorum = 80%
            556 motivos encontrados
      Quorum = 75%
            556 motivos encontrados
      Quorum = 60%
            927 motivos encontrados
Erro = 1
      Quorum = 100%
            7881 motivos encontrados
            k = 1 posicao = 650     chi2 = 2.52 (GCGGGG)
            k = 2 posicao = 773     chi2 = 1.73 (GCGGGG)
            k = 3 posicao = 1341    chi2 = 1.01 (GCGGGG)
            k = 4 posicao = 4437    chi2 = 0.52 (GCGGGG)
      Quorum = 80%
            14601 motivos encontrados
            k = 1 posicao = 1895    chi2 = 2.49 (GCGGGG)
            k = 2 posicao = 2045    chi2 = 1.61 (GCGGGG)
            k = 3 posicao = 3284    chi2 = 1.1  (GCGGGG)
            k = 4 posicao = 7272    chi2 = 0.5  (GCGGGG)
      Quorum = 75%
```

```
            14601 motivos encontrados
            k = 1 posicao = 2121    chi2 = 2.34 (GCGGGG)
            k = 2 posicao = 2604    chi2 = 1.53 (GCGGGG)
            k = 3 posicao = 3778    chi2 = 0.89 (GCGGGG)
            k = 4 posicao = 8057    chi2 = 0.47 (GCGGGG)
      Quorum = 60%
            25117 motivos encontrados
            k = 1 posicao = 4371    chi2 = 2.34 (GCGGGG)
            k = 2 posicao = 5452    chi2 = 1.53 (GCGGGG)
            k = 3 posicao = 7361    chi2 = 1.16 (GCGGGG)
            k = 4 posicao = 12643   chi2 = 0.51 (GCGGGG)


6_0_UME6

Motivo: GCCGCC
Frequência: 3/7 ~42,86%
Tamanho das Sequências: 1000

Erro = 0
      Quorum = 100%
            257 motivos encontrados
      Quorum = 80%
            454 motivos encontrados
      Quorum = 75%
            454 motivos encontrados
      Quorum = 60%
            748 motivos encontrados
Erro = 1
      Quorum = 100%
            7186 motivos encontrados
      Quorum = 80%
            12323 motivos encontrados
      Quorum = 75%
            12323 motivos encontrados
      Quorum = 60%
            19726 motivos encontrados
            k = 1 posicao = 13214   chi2 = 0.02 (GCCGCC)
            k = 2 posicao = 14628   chi2 = 0    (GCCGCC)
            k = 3 posicao = 14684   chi2 = 0    (GCCGCC)
            k = 4 posicao = 15398   chi2 = 0.02 (GCCGCC)


6_1_Gln3

Motivo: GATWAG
Frequência: 11/13 ~84,62%
Tamanho das Sequências: 1000

Erro = 0
      Quorum = 100%
            173 motivos encontrados
      Quorum = 80%
            403 motivos encontrados
      Quorum = 75%
            530 motivos encontrados
            k = 1 posicao = 6       chi2 = 7.55 (GATAAG)
```

```
            k = 2 posicao = 3       chi2 = 8.92 (GATAAG)
            k = 3 posicao = 3       chi2 = 7.25 (GATAAG)
            k = 4 posicao = 12      chi2 = 4.57 (GATAAG)
      Quorum = 60%
            836 motivos encontrados
            k = 1 posicao = 24      chi2 = 7.71 (GATAAG)
            k = 2 posicao = 7       chi2 = 9.4  (GATAAG)
            k = 3 posicao = 12      chi2 = 7.19 (GATAAG)
            k = 4 posicao = 25      chi2 = 4.8  (GATAAG)
Erro = 1
      Quorum = 100%
            5360 motivos encontrados
            k = 1 posicao = 2690    chi2 = 0.12 (GATAAG)
            k = 1 posicao = 3118    chi2 = 0.08 (GATTAG)
            k = 2 posicao = 2917    chi2 = 0.1  (GATAAG)
            k = 2 posicao = 2424    chi2 = 0.2  (GATTAG)
            k = 3 posicao = 3096    chi2 = 0.07 (GATAAG)
            k = 3 posicao = 1982    chi2 = 0.33 (GATTAG)
            k = 4 posicao = 3004    chi2 = 0.07 (GATAAG)
            k = 4 posicao = 2767    chi2 = 0.1  (GATTAG)
      Quorum = 80%
            10819 motivos encontrados
            k = 1 posicao = 6729    chi2 = 0.11 (GATAAG)
            k = 1 posicao = 6816    chi2 = 0.09 (GATTAG)
            k = 2 posicao = 6240    chi2 = 0.13 (GATAAG)
            k = 2 posicao = 6446    chi2 = 0.11 (GATTAG)
            k = 3 posicao = 7301    chi2 = 0.04 (GATAAG)
            k = 3 posicao = 4235    chi2 = 0.53 (GATTAG)
            k = 4 posicao = 7045    chi2 = 0.04 (GATAAG)
            k = 4 posicao = 6317    chi2 = 0.12 (GATTAG)
      Quorum = 75%
            13889 motivos encontrados
            k = 1 posicao = 9061    chi2 = 0.07 (GATAAG)
            k = 1 posicao = 8672    chi2 = 0.1  (GATTAG)
            k = 2 posicao = 8095    chi2 = 0.16 (GATAAG)
            k = 2 posicao = 8763    chi2 = 0.1  (GATTAG)
            k = 3 posicao = 8965    chi2 = 0.08 (GATAAG)
            k = 3 posicao = 6907    chi2 = 0.36 (GATTAG)
            k = 4 posicao = 8966    chi2 = 0.06 (GATAAG)
            k = 4 posicao = 8323    chi2 = 0.11 (GATTAG)
      Quorum = 60%
            22209 motivos encontrados
            k = 1 posicao = 14904   chi2 = 0.08 (GATAAG)
            k = 1 posicao = 14384   chi2 = 0.13 (GATTAG)
            k = 2 posicao = 13923   chi2 = 0.15 (GATAAG)
            k = 2 posicao = 13902   chi2 = 0.15 (GATTAG)
            k = 3 posicao = 14932   chi2 = 0.08 (GATAAG)
            k = 3 posicao = 11534   chi2 = 0.46 (GATTAG)
            k = 4 posicao = 14680   chi2 = 0.07 (GATAAG)
            k = 4 posicao = 13878   chi2 = 0.17 (GATTAG)

6_1_RAP1

Motivo: CACCCR
Frequência: 11/18 ~61,1%
Tamanho das Sequências: 1000
```

```
Erro = 0
      Quorum = 100%
            128 motivos encontrados
      Quorum = 80%
            372 motivos encontrados
      Quorum = 75%
            454 motivos encontrados
      Quorum = 60%
            785 motivos encontrados
Erro = 1
      Quorum = 100%
            4212 motivos encontrados
            k = 1 posicao = 326     chi2 = 1.4  (CACCCA)
            k = 2 posicao = 594     chi2 = 1.08 (CACCCA)
            k = 3 posicao = 244     chi2 = 1.07 (CACCCA)
            k = 4 posicao = 1210    chi2 = 0.38 (CACCCA)
      Quorum = 80%
            9800 motivos encontrados
            k = 1 posicao = 1302    chi2 = 1.57 (CACCCA)
            k = 1 posicao = 2061    chi2 = 0.87 (CACCCG)
            k = 2 posicao = 2058    chi2 = 1.17 (CACCCA)
            k = 2 posicao = 2522    chi2 = 0.59 (CACCCG)
            k = 3 posicao = 1624    chi2 = 1.06 (CACCCA)
            k = 3 posicao = 2994    chi2 = 0.45 (CACCCG)
            k = 4 posicao = 2703    chi2 = 0.5  (CACCCA)
            k = 4 posicao = 3926    chi2 = 0.11 (CACCCG)
      Quorum = 75%
            11933 motivos encontrados
            k = 1 posicao = 1951    chi2 = 1.45 (CACCCA)
            k = 1 posicao = 2903    chi2 = 0.81 (CACCCG)
            k = 2 posicao = 2227    chi2 = 1.16 (CACCCA)
            k = 2 posicao = 3417    chi2 = 0.62 (CACCCG)
            k = 3 posicao = 2137    chi2 = 1.08 (CACCCA)
            k = 3 posicao = 3426    chi2 = 0.41 (CACCCG)
            k = 4 posicao = 4518    chi2 = 0.31 (CACCCA)
            k = 4 posicao = 4805    chi2 = 0.12 (CACCCG)
      Quorum = 60%
            20531 motivos encontrados
            k = 1 posicao = 5137    chi2 = 1.44 (CACCCA)
            k = 1 posicao = 5556    chi2 = 0.73 (CACCCG)
            k = 2 posicao = 4865    chi2 = 1.13 (CACCCA)
            k = 2 posicao = 5737    chi2 = 0.62 (CACCCG)
            k = 3 posicao = 4491    chi2 = 0.97 (CACCCA)
            k = 3 posicao = 5797    chi2 = 0.52 (CACCCG)
            k = 4 posicao = 7825    chi2 = 0.41 (CACCCA)
            k = 4 posicao = 7275    chi2 = 0.18 (CACCCG)

7_0_REB1

Motivo: TTACCC
Frequência: 7/8 ~87,5%
Tamanho das Sequências: 1000

Erro = 0
      Quorum = 100%
            211 motivos encontrados
            k = 1 posicao = 4 chi2 = 9.65  (TTACCC)
```

```
            k = 2 posicao = 4 chi2 = 10.5  (TTACCC)
            k = 3 posicao = 4 chi2 = 10.75 (TTACCC)
            k = 4 posicao = 4 chi2 = 7.55  (TTACCC)
      Quorum = 80%
            405 motivos encontrados
            k = 1 posicao = 4 chi2 = 9.32  (TTACCC)
            k = 2 posicao = 4 chi2 = 10.83 (TTACCC)
            k = 3 posicao = 4 chi2 = 11.06 (TTACCC)
            k = 4 posicao = 4 chi2 = 7.23  (TTACCC)
      Quorum = 75%
            605 motivos encontrados
            k = 1 posicao = 6 chi2 = 9.47  (TTACCC)
            k = 2 posicao = 5 chi2 = 9.99  (TTACCC)
            k = 3 posicao = 5 chi2 = 10.98 (TTACCC)
            k = 4 posicao = 5 chi2 = 8.08  (TTACCC)
      Quorum = 60%
            907 motivos encontrados
            k = 1 posicao = 9 chi2 = 10.07 (TTACCC)
            k = 2 posicao = 6 chi2 = 10.89 (TTACCC)
            k = 3 posicao = 6 chi2 = 10.81 (TTACCC)
            k = 4 posicao = 6 chi2 = 7.88  (TTACCC)
Erro = 1
      Quorum = 100%
            6121 motivos encontrados
            k = 1 posicao = 4617    chi2 = 0.13 (TTACCC)
            k = 2 posicao = 4454    chi2 = 0.16 (TTACCC)
            k = 3 posicao = 4730    chi2 = 0.12 (TTACCC)
            k = 4 posicao = 5483    chi2 = 0.04 (TTACCC)
      Quorum = 80%
            10337 motivos encontrados
            k = 1 posicao = 7525    chi2 = 0.08 (TTACCC)
            k = 2 posicao = 7267    chi2 = 0.1  (TTACCC)
            k = 3 posicao = 7129    chi2 = 0.12 (TTACCC)
            k = 4 posicao = 7469    chi2 = 0.09 (TTACCC)
      Quorum = 75%
            15777 motivos encontrados
            k = 1 posicao = 10287   chi2 = 0.12 (TTACCC)
            k = 2 posicao = 10740   chi2 = 0.1  (TTACCC)
            k = 3 posicao = 9696    chi2 = 0.16 (TTACCC)
            k = 4 posicao = 10789   chi2 = 0.08 (TTACCC)
      Quorum = 60%
            23856 motivos encontrados
            k = 1 posicao = 15646   chi2 = 0.1  (TTACCC)
            k = 2 posicao = 15437   chi2 = 0.12 (TTACCC)
            k = 3 posicao = 13932   chi2 = 0.22 (TTACCC)
            k = 4 posicao = 15597   chi2 = 0.09 (TTACCC)
7_0_RIM101

Motivo: TGCCAAG
Frequência: 5/7 ~71,43%
Tamanho das Sequências: 1000

Erro = 0
      Quorum = 100%
            235 motivos encontrados
      Quorum = 80%
            475 motivos encontrados
```

```
            k = 1 posicao = 3 chi2 = 9.45 (TGCCAAG)
            k = 2 posicao = 2 chi2 = 9.2  (TGCCAAG)
            k = 3 posicao = 2 chi2 = 8.92 (TGCCAAG)
            k = 4 posicao = 3 chi2 = 7    (TGCCAAG)
      Quorum = 75%
            475 motivos encontrados
            k = 1 posicao = 2 chi2 = 9.24 (TGCCAAG)
            k = 2 posicao = 3 chi2 = 8.86 (TGCCAAG)
            k = 3 posicao = 3 chi2 = 8.69 (TGCCAAG)
            k = 4 posicao = 2 chi2 = 6.78 (TGCCAAG)
      Quorum = 60%
            785 motivos encontrados
            k = 1 posicao = 5 chi2 = 9.52 (TGCCAAG)
            k = 2 posicao = 4 chi2 = 9.34 (TGCCAAG)
            k = 3 posicao = 4 chi2 = 8.69 (TGCCAAG)
            k = 4 posicao = 6 chi2 = 6.7  (TGCCAAG)
Erro = 1
      Quorum = 100%
            7028 motivos encontrados
            k = 1 posicao = 436     chi2 = 3.21 (TGCCAAG)
            k = 2 posicao = 551     chi2 = 2.17 (TGCCAAG)
            k = 3 posicao = 940     chi2 = 2.23 (TGCCAAG)
            k = 4 posicao = 1492    chi2 = 0.95 (TGCCAAG)
      Quorum = 80%
            12574 motivos encontrados
            k = 1 posicao = 1703    chi2 = 3.25 (TGCCAAG)
            k = 2 posicao = 2035    chi2 = 2.44 (TGCCAAG)
            k = 3 posicao = 1678    chi2 = 2.17 (TGCCAAG)
            k = 4 posicao = 3298    chi2 = 0.93 (TGCCAAG)
      Quorum = 75%
            12574 motivos encontrados
            k = 1 posicao = 1739    chi2 = 3.33 (TGCCAAG)
            k = 2 posicao = 1717    chi2 = 2.16 (TGCCAAG)
            k = 3 posicao = 1609    chi2 = 2.12 (TGCCAAG)
            k = 4 posicao = 3574    chi2 = 0.75 (TGCCAAG)
      Quorum = 60%
            20257 motivos encontrados
            k = 1 posicao = 3598    chi2 = 3.04 (TGCCAAG)
            k = 2 posicao = 3664    chi2 = 2.11 (TGCCAAG)
            k = 3 posicao = 3206    chi2 = 2.13 (TGCCAAG)
            k = 4 posicao = 4753    chi2 = 1.15 (TGCCAAG)



7_0_STE12

Motivo: TGAAACA
Frequência: 6/6 ~100%
Tamanho das Sequências: 1000

Erro = 0
      Quorum = 100%
            267 motivos encontrados
            k = 1 posicao = 4 chi2 = 9.98 (TGAAACA)
            k = 2 posicao = 5 chi2 = 9.08 (TGAAACA)
            k = 3 posicao = 5 chi2 = 8.75 (TGAAACA)
```

```
            k = 4 posicao = 4 chi2 = 7.43 (TGAAACA)
      Quorum = 80%
            529 motivos encontrados
            k = 1 posicao = 4 chi2 = 9.69 (TGAAACA)
            k = 2 posicao = 6 chi2 = 8.4  (TGAAACA)
            k = 3 posicao = 5 chi2 = 8.45 (TGAAACA)
            k = 4 posicao = 5 chi2 = 7.02 (TGAAACA)
      Quorum = 75%
            529 motivos encontrados
            k = 1 posicao = 6 chi2 = 9.33 (TGAAACA)
            k = 2 posicao = 5 chi2 = 8.66 (TGAAACA)
            k = 3 posicao = 5 chi2 = 8.6  (TGAAACA)
            k = 4 posicao = 5 chi2 = 7.17 (TGAAACA)
      Quorum = 60%
            916 motivos encontrados
            k = 1 posicao = 12      chi2 = 9.56 (TGAAACA)
            k = 2 posicao = 13      chi2 = 8.93 (TGAAACA)
            k = 3 posicao = 16      chi2 = 8.4  (TGAAACA)
            k = 4 posicao = 18      chi2 = 7.15 (TGAAACA)
Erro = 1
      Quorum = 100%
            7705 motivos encontrados
            k = 1 posicao = 4203    chi2 = 0.55 (TGAAACA)
            k = 2 posicao = 5562    chi2 = 0.2  (TGAAACA)
            k = 3 posicao = 5481    chi2 = 0.21 (TGAAACA)
            k = 4 posicao = 6219    chi2 = 0.12 (TGAAACA)
      Quorum = 80%
            14161 motivos encontrados
            k = 1 posicao = 8113    chi2 = 0.52 (TGAAACA)
            k = 2 posicao = 9261    chi2 = 0.22 (TGAAACA)
            k = 3 posicao = 8536    chi2 = 0.29 (TGAAACA)
            k = 4 posicao = 10343   chi2 = 0.1 (TGAAACA)
      Quorum = 75%
            14161 motivos encontrados
            k = 1 posicao = 7156    chi2 = 0.66 (TGAAACA)
            k = 2 posicao = 8778    chi2 = 0.28 (TGAAACA)
            k = 3 posicao = 8848    chi2 = 0.34 (TGAAACA)
            k = 4 posicao = 9747    chi2 = 0.15 (TGAAACA)
      Quorum = 60%
            24421 motivos encontrados
            k = 1 posicao = 12918   chi2 = 0.49 (TGAAACA)
            k = 2 posicao = 15781   chi2 = 0.24 (TGAAACA)
            k = 3 posicao = 13941   chi2 = 0.3  (TGAAACA)
            k = 4 posicao = 16147   chi2 = 0.17 (TGAAACA)
7_1_SWI4

Motivo: CRCGAAA
Frequência: 5/5 ~100%
Tamanho das Sequências: 1000

Erro = 0
      Quorum = 100%
            330 motivos encontrados
      Quorum = 80%
            675 motivos encontrados
            k = 1 posicao = 3       chi2 = 5.75 (CGCGAAA)
            k = 2 posicao = 3       chi2 = 4.83 (CGCGAAA)
```

```
            k = 3 posicao = 11      chi2 = 3.95 (CGCGAAA)
            k = 4 posicao = 5       chi2 = 3.67 (CGCGAAA)
      Quorum = 75%
            675 motivos encontrados
            k = 1 posicao = 2       chi2 = 6    (CGCGAAA)
            k = 2 posicao = 4       chi2 = 5.04 (CGCGAAA)
            k = 3 posicao = 14      chi2 = 3.72 (CGCGAAA)
            k = 4 posicao = 14      chi2 = 3.43 (CGCGAAA)
      Quorum = 60%
            1349 motivos encontrados
            k = 1 posicao = 50      chi2 = 3.1  (CACGAAA)
            k = 1 posicao = 8       chi2 = 5.69 (CGCGAAA)
            k = 2 posicao = 81      chi2 = 2.6  (CACGAAA)
            k = 2 posicao = 20      chi2 = 4.8  (CGCGAAA)
            k = 3 posicao = 117     chi2 = 2.1  (CACGAAA)
            k = 3 posicao = 47      chi2 = 3.7  (CGCGAAA)
            k = 4 posicao = 309     chi2 = 0.56 (CACGAAA)
            k = 4 posicao = 33      chi2 = 3.67 (CGCGAAA)
Erro = 1
      Quorum = 100%
            8974 motivos encontrados
            k = 1 posicao = 2630    chi2 = 0.93 (CACGAAA)
            k = 1 posicao = 1189    chi2 = 2.12 (CGCGAAA)
            k = 2 posicao = 3046    chi2 = 0.7  (CACGAAA)
            k = 2 posicao = 2138    chi2 = 1.01 (CGCGAAA)
            k = 3 posicao = 4224    chi2 = 0.3  (CACGAAA)
            k = 3 posicao = 4064    chi2 = 0.4  (CGCGAAA)
            k = 4 posicao = 5230    chi2 = 0.12 (CACGAAA)
            k = 4 posicao = 4771    chi2 = 0.18 (CGCGAAA)
      Quorum = 80%
            17937 motivos encontrados
            k = 1 posicao = 6152    chi2 = 1.07 (CACGAAA)
            k = 1 posicao = 3575    chi2 = 2.11 (CGCGAAA)
            k = 2 posicao = 7073    chi2 = 0.71 (CACGAAA)
            k = 2 posicao = 6979    chi2 = 0.78 (CGCGAAA)
            k = 3 posicao = 9212    chi2 = 0.35 (CACGAAA)
            k = 3 posicao = 8502    chi2 = 0.38 (CGCGAAA)
            k = 4 posicao = 10421   chi2 = 0.17 (CACGAAA)
            k = 4 posicao = 9629    chi2 = 0.26 (CGCGAAA)
      Quorum = 75%
            17937 motivos encontrados
            k = 1 posicao = 6271    chi2 = 1.04 (CACGAAA)
            k = 1 posicao = 4092    chi2 = 1.71 (CGCGAAA)
            k = 2 posicao = 8574    chi2 = 0.52 (CACGAAA)
            k = 2 posicao = 5741    chi2 = 1.07 (CGCGAAA)
            k = 3 posicao = 10092   chi2 = 0.26 (CACGAAA)
            k = 3 posicao = 8743    chi2 = 0.44 (CGCGAAA)
            k = 4 posicao = 11741   chi2 = 0.09 (CACGAAA)
            k = 4 posicao = 9601    chi2 = 0.22 (CGCGAAA)
      Quorum = 60%
            35352 motivos encontrados
            k = 1 posicao = 15854   chi2 = 0.94 (CACGAAA)
            k = 1 posicao = 9842    chi2 = 1.96 (CGCGAAA)
            k = 2 posicao = 18655   chi2 = 0.63 (CACGAAA)
            k = 2 posicao = 16854   chi2 = 0.93 (CGCGAAA)
            k = 3 posicao = 21349   chi2 = 0.25 (CACGAAA)
            k = 3 posicao = 19789   chi2 = 0.36 (CGCGAAA)
```

```
            k = 4 posicao = 22509   chi2 = 0.16 (CACGAAA)
            k = 4 posicao = 21049   chi2 = 0.27 (CGCGAAA)


7_2_GCN4

Motivo: TGASTCW
Frequência: 0/9 ~0%
Tamanho das Sequências: 1000

Erro = 0
      Quorum = 100%
            229 motivos encontrados
      Quorum = 80%
            372 motivos encontrados
      Quorum = 75%
            586 motivos encontrados
      Quorum = 60%
            839 motivos encontrados
Erro = 1
      Quorum = 100%
            6343 motivos encontrados
      Quorum = 80%
            10422 motivos encontrados
      Quorum = 75%
            15227 motivos encontrados
      Quorum = 60%
            21609 motivos encontrados



7_2_RGT1

Motivo: CGGADDA
Frequência: 10/10 ~100%
Tamanho das Sequências: 1000

Erro = 0
      Quorum = 100%
            205 motivos encontrados
      Quorum = 80%
            496 motivos encontrados
            k = 1 posicao = 4       chi2 = 10.52 (CGGAAAA)
            k = 2 posicao = 2       chi2 = 8.83  (CGGAAAA)
            k = 3 posicao = 2       chi2 = 6.92  (CGGAAAA)
            k = 4 posicao = 10      chi2 = 3.08  (CGGAAAA)
      Quorum = 75%
            496 motivos encontrados
            k = 1 posicao = 4 chi2 = 11.01 (CGGAAAA)
            k = 2 posicao = 3 chi2 = 8.55  (CGGAAAA)
            k = 3 posicao = 4 chi2 = 6.85  (CGGAAAA)
            k = 4 posicao = 9 chi2 = 3.38  (CGGAAAA)
      Quorum = 60%
            987 motivos encontrados
            k = 1 posicao = 10      chi2 = 10.87 (CGGAAAA)
            k = 2 posicao = 11      chi2 = 7.92  (CGGAAAA)
            k = 3 posicao = 10      chi2 = 6.37  (CGGAAAA)
```

```
            k = 4 posicao = 27      chi2 = 3.33  (CGGAAAA)
Erro = 1
      Quorum = 100%
            6085 motivos encontrados
            k = 1 posicao = 740     chi2 = 2.26 (CGGAAAA)
            k = 1 posicao = 244     chi2 = 4.42 (CGGAAGA)
            k = 1 posicao = 375     chi2 = 4.15 (CGGAGAA)
            k = 1 posicao = 795     chi2 = 2.05 (CGGATAA)
            k = 2 posicao = 1431    chi2 = 0.69 (CGGAAAA)
            k = 2 posicao = 350     chi2 = 2.57 (CGGAAGA)
            k = 2 posicao = 313     chi2 = 2.81 (CGGAGAA)
            k = 2 posicao = 403     chi2 = 2.57 (CGGATAA)
            k = 3 posicao = 2933    chi2 = 0.15 (CGGAAAA)
            k = 3 posicao = 543     chi2 = 1.61 (CGGAAGA)
            k = 3 posicao = 758     chi2 = 1.62 (CGGAGAA)
            k = 3 posicao = 334     chi2 = 2.42 (CGGATAA)
            k = 4 posicao = 3388    chi2 = 0.06 (CGGAAAA)
            k = 4 posicao = 944     chi2 = 0.88 (CGGAAGA)
            k = 4 posicao = 682     chi2 = 0.96 (CGGAGAA)
            k = 4 posicao = 448     chi2 = 1.57 (CGGATAA)
      Quorum = 80%
            13578 motivos encontrados
            k = 1 posicao = 2782    chi2 = 2.11 (CGGAAAA)
            k = 1 posicao = 1309    chi2 = 4.43 (CGGAAGA)
            k = 1 posicao = 9793    chi2 = 0    (CGGAATA)
            k = 1 posicao = 1525    chi2 = 4.36 (CGGAGAA)
            k = 1 posicao = 1267    chi2 = 3.51 (CGGAGGA)
            k = 1 posicao = 2001    chi2 = 2.82 (CGGATAA)
            k = 1 posicao = 4454    chi2 = 0.55 (CGGATTA)
            k = 2 posicao = 4354    chi2 = 0.63 (CGGAAAA)
            k = 2 posicao = 1547    chi2 = 2.73 (CGGAAGA)
            k = 2 posicao = 7920    chi2 = 0.04 (CGGAATA)
            k = 2 posicao = 2260    chi2 = 2.62 (CGGAGAA)
            k = 2 posicao = 1049    chi2 = 2.71 (CGGAGGA)
            k = 2 posicao = 1502    chi2 = 2.71 (CGGATAA)
            k = 2 posicao = 3568    chi2 = 0.83 (CGGATTA)
            k = 3 posicao = 6889    chi2 = 0.2  (CGGAAAA)
            k = 3 posicao = 2122    chi2 = 1.6  (CGGAAGA)
            k = 3 posicao = 10332   chi2 = 0.1  (CGGAATA)
            k = 3 posicao = 2408    chi2 = 1.44 (CGGAGAA)
            k = 3 posicao = 1273    chi2 = 2.14 (CGGAGGA)
            k = 3 posicao = 1406    chi2 = 2.11 (CGGATAA)
            k = 3 posicao = 1203    chi2 = 1.89 (CGGATTA)
            k = 4 posicao = 8380    chi2 = 0.03 (CGGAAAA)
            k = 4 posicao = 4250    chi2 = 0.74 (CGGAAGA)
            k = 4 posicao = 12881   chi2 = 0.37 (CGGAATA)
            k = 4 posicao = 3392    chi2 = 0.88 (CGGAGAA)
            k = 4 posicao = 1278    chi2 = 1.62 (CGGAGGA)
            k = 4 posicao = 1930    chi2 = 1.48 (CGGATAA)
            k = 4 posicao = 1157    chi2 = 1.36 (CGGATTA)
      Quorum = 75%
            13578 motivos encontrados
            k = 1 posicao = 2620    chi2 = 2.64 (CGGAAAA)
            k = 1 posicao = 1452    chi2 = 4.43 (CGGAAGA)
            k = 1 posicao = 8943    chi2 = 0.01 (CGGAATA)
            k = 1 posicao = 1121    chi2 = 4.46 (CGGAGAA)
            k = 1 posicao = 965     chi2 = 4.24 (CGGAGGA)
```

```
        k = 1 posicao = 2453    chi2 = 2.46 (CGGATAA)
        k = 1 posicao = 3571    chi2 = 0.86 (CGGATTA)
        k = 2 posicao = 4690    chi2 = 0.69 (CGGAAAA)
        k = 2 posicao = 1475    chi2 = 2.55 (CGGAAGA)
        k = 2 posicao = 7150    chi2 = 0.1  (CGGAATA)
        k = 2 posicao = 1209    chi2 = 2.59 (CGGAGAA)
        k = 2 posicao = 837     chi2 = 2.74 (CGGAGGA)
        k = 2 posicao = 1900    chi2 = 2.68 (CGGATAA)
        k = 2 posicao = 3981    chi2 = 0.66 (CGGATTA)
        k = 3 posicao = 6549    chi2 = 0.22 (CGGAAAA)
        k = 3 posicao = 1447    chi2 = 1.91 (CGGAAGA)
        k = 3 posicao = 10656   chi2 = 0.18 (CGGAATA)
        k = 3 posicao = 2435    chi2 = 1.55 (CGGAGAA)
        k = 3 posicao = 1026    chi2 = 2.1  (CGGAGGA)
        k = 3 posicao = 1469    chi2 = 2.33 (CGGATAA)
        k = 3 posicao = 1202    chi2 = 1.89 (CGGATTA)
        k = 4 posicao = 11692   chi2 = 0    (CGGAAAA)
        k = 4 posicao = 3717    chi2 = 0.77 (CGGAAGA)
        k = 4 posicao = 10890   chi2 = 0.35 (CGGAATA)
        k = 4 posicao = 3482    chi2 = 0.87 (CGGAGAA)
        k = 4 posicao = 1187    chi2 = 1.68 (CGGAGGA)
        k = 4 posicao = 1746    chi2 = 1.48 (CGGATAA)
        k = 4 posicao = 1204    chi2 = 1.49 (CGGATTA)
Quorum = 60%
        25839 motivos encontrados
        k = 1 posicao = 5870    chi2 = 2.47 (CGGAAAA)
        k = 1 posicao = 4726    chi2 = 4.05 (CGGAAGA)
        k = 1 posicao = 18076   chi2 = 0    (CGGAATA)
        k = 1 posicao = 3940    chi2 = 3.88 (CGGAGAA)
        k = 1 posicao = 3424    chi2 = 3.84 (CGGAGGA)
        k = 1 posicao = 15679   chi2 = 0.05 (CGGAGTA)
        k = 1 posicao = 6177    chi2 = 2.45 (CGGATAA)
        k = 1 posicao = 15167   chi2 = 0.08 (CGGATGA)
        k = 1 posicao = 11165   chi2 = 0.59 (CGGATTA)
        k = 2 posicao = 11428   chi2 = 0.46 (CGGAAAA)
        k = 2 posicao = 4811    chi2 = 2.42 (CGGAAGA)
        k = 2 posicao = 16216   chi2 = 0.03 (CGGAATA)
        k = 2 posicao = 4128    chi2 = 2.26 (CGGAGAA)
        k = 2 posicao = 3005    chi2 = 2.78 (CGGAGGA)
        k = 2 posicao = 10245   chi2 = 0.4  (CGGAGTA)
        k = 2 posicao = 3437    chi2 = 3    (CGGATAA)
        k = 2 posicao = 10379   chi2 = 0.33 (CGGATGA)
        k = 2 posicao = 8888    chi2 = 0.72 (CGGATTA)
        k = 3 posicao = 13836   chi2 = 0.17 (CGGAAAA)
        k = 3 posicao = 4943    chi2 = 1.91 (CGGAAGA)
        k = 3 posicao = 20550   chi2 = 0.1  (CGGAATA)
        k = 3 posicao = 6774    chi2 = 1.55 (CGGAGAA)
        k = 3 posicao = 2524    chi2 = 2.22 (CGGAGGA)
        k = 3 posicao = 13828   chi2 = 0.1  (CGGAGTA)
        k = 3 posicao = 4859    chi2 = 1.97 (CGGATAA)
        k = 3 posicao = 9284    chi2 = 0.29 (CGGATGA)
        k = 3 posicao = 3146    chi2 = 1.79 (CGGATTA)
        k = 4 posicao = 17095   chi2 = 0.01 (CGGAAAA)
        k = 4 posicao = 7930    chi2 = 0.83 (CGGAAGA)
        k = 4 posicao = 24667   chi2 = 0.51 (CGGAATA)
        k = 4 posicao = 9042    chi2 = 0.83 (CGGAGAA)
        k = 4 posicao = 3884    chi2 = 1.47 (CGGAGGA)
```

```
            k = 4 posicao = 13905   chi2 = 0.07 (CGGAGTA)
            k = 4 posicao = 4859    chi2 = 1.56 (CGGATAA)
            k = 4 posicao = 14606   chi2 = 0.05 (CGGATGA)
            k = 4 posicao = 3783    chi2 = 1.27 (CGGATTA)


8_1_PDR1

Motivo: TCCGYGGA
Frequência: 6/8 ~75%
Tamanho das Sequências: 1000

Erro = 0
      Quorum = 100%
            229 motivos encontrados
      Quorum = 80%
            416 motivos encontrados
      Quorum = 75%
            639 motivos encontrados
            k = 1 posicao = 2 chi2 = 9.6  (TCCGCGGA)
            k = 2 posicao = 2 chi2 = 9.41 (TCCGCGGA)
            k = 3 posicao = 2 chi2 = 9.15 (TCCGCGGA)
            k = 4 posicao = 2 chi2 = 6.74 (TCCGCGGA)
      Quorum = 60%
            950 motivos encontrados
            k = 1 posicao = 4 chi2 = 9.37 (TCCGCGGA)
            k = 2 posicao = 3 chi2 = 9.33 (TCCGCGGA)
            k = 3 posicao = 3 chi2 = 9.26 (TCCGCGGA)
            k = 4 posicao = 3 chi2 = 6.86 (TCCGCGGA)
Erro = 1
      Quorum = 100%
            6610 motivos encontrados
            k = 1 posicao = 8       chi2 = 10.95 (TCCGCGGA)
            k = 1 posicao = 15      chi2 = 10.07 (TCCGTGGA)
            k = 2 posicao = 8       chi2 = 10.92 (TCCGCGGA)
            k = 2 posicao = 18      chi2 = 10.39 (TCCGTGGA)
            k = 3 posicao = 11      chi2 = 9.07  (TCCGCGGA)
            k = 3 posicao = 9       chi2 = 8.81  (TCCGTGGA)
            k = 4 posicao = 18      chi2 = 5.17  (TCCGCGGA)
            k = 4 posicao = 8       chi2 = 7.59  (TCCGTGGA)
      Quorum = 80%
            11058 motivos encontrados
            k = 1 posicao = 32      chi2 = 11.26 (TCCGCGGA)
            k = 1 posicao = 48      chi2 = 10.53 (TCCGTGGA)
            k = 2 posicao = 21      chi2 = 10.53 (TCCGCGGA)
            k = 2 posicao = 19      chi2 = 10.98 (TCCGTGGA)
            k = 3 posicao = 29      chi2 = 8.81 (TCCGCGGA)
            k = 3 posicao = 60      chi2 = 8.08 (TCCGTGGA)
            k = 4 posicao = 97      chi2 = 5.09 (TCCGCGGA)
            k = 4 posicao = 42      chi2 = 6.98 (TCCGTGGA)
      Quorum = 75%
            16695 motivos encontrados
            k = 1 posicao = 92      chi2 = 11.26 (TCCGCGGA)
            k = 1 posicao = 178     chi2 = 10.26 (TCCGTGGA)
            k = 2 posicao = 70      chi2 = 10.92 (TCCGCGGA)
            k = 2 posicao = 73      chi2 = 10.69 (TCCGTGGA)
            k = 3 posicao = 64      chi2 = 9.83 (TCCGCGGA)
```

```
            k = 3 posicao = 83      chi2 = 8.98 (TCCGTGGA)
            k = 4 posicao = 336     chi2 = 4.8 (TCCGCGGA)
            k = 4 posicao = 130     chi2 = 6.57 (TCCGTGGA)
      Quorum = 60%
            25049 motivos encontrados
            k = 1 posicao = 220     chi2 = 11.35 (TCCGCGGA)
            k = 1 posicao = 369     chi2 = 10.04 (TCCGTGGA)
            k = 2 posicao = 83      chi2 = 11.59 (TCCGCGGA)
            k = 2 posicao = 263     chi2 = 10.45 (TCCGTGGA)
            k = 3 posicao = 134     chi2 = 8.83 (TCCGCGGA)
            k = 3 posicao = 206     chi2 = 8.57 (TCCGTGGA)
            k = 4 posicao = 580     chi2 = 4.88 (TCCGCGGA)
            k = 4 posicao = 213     chi2 = 6.7 (TCCGTGGA)



8_1_PDR3

Motivo: TCCGYGGA
Frequência: 7/8 ~87,5%
Tamanho das Sequências: 1000

Erro = 0
      Quorum = 100%
            241 motivos encontrados
      Quorum = 80%
            423 motivos encontrados
            k = 1 posicao = 2 chi2 = 12.37 (TCCGCGGA)
            k = 2 posicao = 2 chi2 = 12.17 (TCCGCGGA)
            k = 3 posicao = 2 chi2 = 12.29 (TCCGCGGA)
            k = 4 posicao = 2 chi2 = 9.15  (TCCGCGGA)
      Quorum = 75%
            667 motivos encontrados
            k = 1 posicao = 2 chi2 = 12.37 (TCCGCGGA)
            k = 2 posicao = 2 chi2 = 12.17 (TCCGCGGA)
            k = 3 posicao = 2 chi2 = 12.17 (TCCGCGGA)
            k = 4 posicao = 2 chi2 = 9.09  (TCCGCGGA)
      Quorum = 60%
            1024 motivos encontrados
            k = 1 posicao = 3 chi2 = 12.33 (TCCGCGGA)
            k = 2 posicao = 4 chi2 = 12.4  (TCCGCGGA)
            k = 3 posicao = 3 chi2 = 12.21 (TCCGCGGA)
            k = 4 posicao = 5 chi2 = 9.18  (TCCGCGGA)
Erro = 1
      Quorum = 100%
            6897 motivos encontrados
            k = 1 posicao = 17      chi2 = 11.15 (TCCGCGGA)
            k = 1 posicao = 24      chi2 = 10.26 (TCCGTGGA)
            k = 2 posicao = 12      chi2 = 11.74 (TCCGCGGA)
            k = 2 posicao = 10      chi2 = 11.38 (TCCGTGGA)
            k = 3 posicao = 17      chi2 = 9.45  (TCCGCGGA)
            k = 3 posicao = 29      chi2 = 8.27  (TCCGTGGA)
            k = 4 posicao = 65      chi2 = 4.81  (TCCGCGGA)
            k = 4 posicao = 39      chi2 = 6.16  (TCCGTGGA)
      Quorum = 80%
            11592 motivos encontrados
            k = 1 posicao = 69      chi2 = 11.56 (TCCGCGGA)
```

```
            k = 1 posicao = 117     chi2 = 10.12 (TCCGTGGA)
            k = 2 posicao = 38      chi2 = 11.77 (TCCGCGGA)
            k = 2 posicao = 69      chi2 = 10.61 (TCCGTGGA)
            k = 3 posicao = 76      chi2 = 8.93  (TCCGCGGA)
            k = 3 posicao = 53      chi2 = 8.17  (TCCGTGGA)
            k = 4 posicao = 131     chi2 = 5.26  (TCCGCGGA)
            k = 4 posicao = 61      chi2 = 6.38  (TCCGTGGA)
      Quorum = 75%
            17555 motivos encontrados
            k = 1 posicao = 210     chi2 = 11.09 (TCCGCGGA)
            k = 1 posicao = 265     chi2 = 10.61 (TCCGTGGA)
            k = 2 posicao = 78      chi2 = 11.5  (TCCGCGGA)
            k = 2 posicao = 102     chi2 = 11.09 (TCCGTGGA)
            k = 3 posicao = 184     chi2 = 9.05  (TCCGCGGA)
            k = 3 posicao = 105     chi2 = 8.5   (TCCGTGGA)
            k = 4 posicao = 424     chi2 = 4.51  (TCCGCGGA)
            k = 4 posicao = 476     chi2 = 5.82  (TCCGTGGA)
      Quorum = 60%
            26303 motivos encontrados
            k = 1 posicao = 433     chi2 = 11.18 (TCCGCGGA)
            k = 1 posicao = 521     chi2 = 9.91  (TCCGTGGA)
            k = 2 posicao = 237     chi2 = 10.64 (TCCGCGGA)
            k = 2 posicao = 322     chi2 = 10.45 (TCCGTGGA)
            k = 3 posicao = 316     chi2 = 9.5   (TCCGCGGA)
            k = 3 posicao = 415     chi2 = 8.27  (TCCGTGGA)
            k = 4 posicao = 978     chi2 = 4.9   (TCCGCGGA)
            k = 4 posicao = 1277    chi2 = 6.16  (TCCGTGGA)


8_1_PDR8

Motivo: TCCGHGGA
Frequência: 3/5 ~60%
Tamanho das Sequências: 1000

Erro = 0
      Quorum = 100%
            328 motivos encontrados
      Quorum = 80%
            685 motivos encontrados
      Quorum = 75%
            685 motivos encontrados
      Quorum = 60%
            1319 motivos encontrados
            k = 1 posicao = 4 chi2 = 4.11 (TCCGTGGA)
            k = 2 posicao = 2 chi2 = 4.21 (TCCGTGGA)
            k = 3 posicao = 2 chi2 = 4.07 (TCCGTGGA)
            k = 4 posicao = 3 chi2 = 3.54 (TCCGTGGA)
Erro = 1
      Quorum = 100%
            9298 motivos encontrados
            k = 1 posicao = 95      chi2 = 6.31 (TCCGAGGA)
            k = 1 posicao = 63      chi2 = 6.75 (TCCGCGGA)
            k = 1 posicao = 75      chi2 = 6.78 (TCCGTGGA)
            k = 2 posicao = 51      chi2 = 6.1  (TCCGAGGA)
            k = 2 posicao = 25      chi2 = 7.45 (TCCGCGGA)
            k = 2 posicao = 22      chi2 = 7.18 (TCCGTGGA)
```

```
      k = 3 posicao = 38      chi2 = 5.38 (TCCGAGGA)
      k = 3 posicao = 68      chi2 = 5.43 (TCCGCGGA)
      k = 3 posicao = 64      chi2 = 5.7  (TCCGTGGA)
      k = 4 posicao = 48      chi2 = 5.29 (TCCGAGGA)
      k = 4 posicao = 161     chi2 = 3.89 (TCCGCGGA)
      k = 4 posicao = 78      chi2 = 4.77 (TCCGTGGA)
Quorum = 80%
      18263 motivos encontrados
      k = 1 posicao = 551     chi2 = 5.92 (TCCGAGGA)
      k = 1 posicao = 147     chi2 = 7.54 (TCCGCGGA)
      k = 1 posicao = 282     chi2 = 6.47 (TCCGTGGA)
      k = 2 posicao = 296     chi2 = 6.05 (TCCGAGGA)
      k = 2 posicao = 143     chi2 = 7.15 (TCCGCGGA)
      k = 2 posicao = 182     chi2 = 6.81 (TCCGTGGA)
      k = 3 posicao = 589     chi2 = 5.11 (TCCGAGGA)
      k = 3 posicao = 347     chi2 = 5.53 (TCCGCGGA)
      k = 3 posicao = 245     chi2 = 5.9  (TCCGTGGA)
      k = 4 posicao = 429     chi2 = 4.62 (TCCGAGGA)
      k = 4 posicao = 944     chi2 = 4.01 (TCCGCGGA)
      k = 4 posicao = 577     chi2 = 4.1  (TCCGTGGA)
Quorum = 75%
      18263 motivos encontrados
      k = 1 posicao = 790     chi2 = 6    (TCCGAGGA)
      k = 1 posicao = 168     chi2 = 7.06 (TCCGCGGA)
      k = 1 posicao = 451     chi2 = 6.31 (TCCGTGGA)
      k = 2 posicao = 195     chi2 = 6.23 (TCCGAGGA)
      k = 2 posicao = 152     chi2 = 6.95 (TCCGCGGA)
      k = 2 posicao = 155     chi2 = 6.86 (TCCGTGGA)
      k = 3 posicao = 159     chi2 = 5.46 (TCCGAGGA)
      k = 3 posicao = 286     chi2 = 5.46 (TCCGCGGA)
      k = 3 posicao = 281     chi2 = 5.85 (TCCGTGGA)
      k = 4 posicao = 342     chi2 = 4.6  (TCCGAGGA)
      k = 4 posicao = 562     chi2 = 3.81 (TCCGCGGA)
      k = 4 posicao = 632     chi2 = 4.01 (TCCGTGGA)
Quorum = 60%
      35494 motivos encontrados
      k = 1 posicao = 1806    chi2 = 5.48 (TCCGAGGA)
      k = 1 posicao = 1580    chi2 = 6.5  (TCCGCGGA)
      k = 1 posicao = 1668    chi2 = 5.95 (TCCGTGGA)
      k = 2 posicao = 1224    chi2 = 5.92 (TCCGAGGA)
      k = 2 posicao = 794     chi2 = 7.01 (TCCGCGGA)
      k = 2 posicao = 991     chi2 = 6.58 (TCCGTGGA)
      k = 3 posicao = 1647    chi2 = 5.36 (TCCGAGGA)
      k = 3 posicao = 2029    chi2 = 5.46 (TCCGCGGA)
      k = 3 posicao = 1118    chi2 = 5.7  (TCCGTGGA)
      k = 4 posicao = 1549    chi2 = 5.02 (TCCGAGGA)
      k = 4 posicao = 2429    chi2 = 3.85 (TCCGCGGA)
      k = 4 posicao = 2813    chi2 = 4.29 (TCCGTGGA)
```

Anexo B Comparação da Significância Estatística referente à Abundância obtida pelo Método de *Shuffling* e pelo Método Analítico

Foi feita uma procura, usando o SMILE, em vários conjuntos de dados.

A parametrização usada fez variar os seguintes parâmetros:

k = 1,...,4
erro = 0,1
quórum = 60, 75, 80, 100

A procura foi feita para motivos com comprimento de 4 a 10, excepto no conjunto de dados ZAP1 onde a procura foi feita para tamanhos entre 4 e 11.

De seguida apresentamos a comparação entre o valor do chi2 obtido pelo método de *shuffling* e o valor obtido pelo método analítico.

Cada linha diz respeito a dados em que se conserva um determinado *k-mer* (*k*). Em cada uma é apresentado o valor do chi2 obtido pelo método de *shuffling* (chi2) e a sua posição relativa na lista de motivos ordenada por valores decrescentes de chi2 face ao número de motivos extraídos. Os valores acima referidos também são apresentados para o método analítico (achi2).

```
10_2_YRR1

Motivo: TTCCGTGGAA   TTCCGCGGAT    TTCCGCGGAA
Tamanho das Sequências: 1000

Error = 0
    Quorum = 100%

Error = 1
    Quorum = 100%
      K=1 chi2=4.88 (#633 in  8908) achi2=4.76 (#745 in 8908) [TTCCGCGGAA]
      K=2 chi2=4.65 (#468 in  8908) achi2=4.64 (#509 in 8908) [TTCCGCGGAA]
      K=3 chi2=4.12 (#201 in 8908) achi2=4.13 (#230 in 8908) [TTCCGCGGAA]
      K=4 chi2=2.53 (#223 in 8908) achi2=2.78 (#205 in 8908) [TTCCGCGGAA]
      K=5 chi2=1.26 (#277 in 8912) achi2=1.43 (#291 in 8912) [TTCCGCGGAA]
      K=6 chi2=0.39 (#323 in 8912) achi2=0.41 (#1035 in 8912) [TTCCGCGGAA]

11_2_ZAP1

Motivo: ACCTTGAAGGT     ACCCTAAAGGT     ACCTTAAAGGT
Frequência: 4/5 ~80%
Tamanho das Sequências: 1000

Error = 0
    Quorum = 80%
```

```
Error = 1
    Quorum = 80%
      K=1 chi2=3.88 (#428 in 17722) achi2=3.87 (#504 in 17722)
[ACCCTAAAGGT]
      K=1 chi2=4.74 (#246 in 17722) achi2=4.83 (#265 in 17722)
[ACCTTAAAGGT]
      K=2 chi2=3.94 (#222 in 17722) achi2=3.89 (#248 in 17722)
[ACCCTAAAGGT]
      K=2 chi2=4.74 (#111 in 17722) achi2=4.84 (#119 in 17722)
[ACCTTAAAGGT]
      K=3 chi2=4.00 (#51 in 17722) achi2=3.91 (#61 in 17722) [ACCCTAAAGGT]
      K=3 chi2=4.77 (#12 in 17722) achi2=4.85 (#14 in 17722) [ACCTTAAAGGT]
      K=4 chi2=3.94 (#28 in 17722) achi2=3.87 (#31 in 17722) [ACCCTAAAGGT]
      K=4 chi2=4.88 (#4 in 17722) achi2=4.85 (#4 in 17722) [ACCTTAAAGGT]
      K=5 chi2=3.26 (#24 in 17722) achi2=3.43 (#29 in 17722) [ACCCTAAAGGT]
      K=5 chi2=4.35 (#7 in 17722) achi2=4.62 (#3 in 17722) [ACCTTAAAGGT]
      K=6 chi2=0.71 (#387 in 17722) achi2=2.82 (#276 in 17722)
[ACCCTAAAGGT]
      K=6 chi2=3.10 (#2 in 17722) achi2=4.31 (#259 in 17722) [ACCTTAAAGGT]


5_0_MSN4

Motivo: CCCCT
Frequência: 3/5 ~60%
Tamanho das Sequências: 1000


Error = 0
    Quorum = 60%
      K=1 chi2=3.44 (#91 in 1316) achi2=3.14 (#116 in 1316) [CCCCT]
      K=2 chi2=1.87 (#183 in 1316) achi2=1.94 (#178 in 1316) [CCCCT]
      K=3 chi2=1.64 (#138 in 1316) achi2=1.40 (#190 in 1316) [CCCCT]
      K=4 chi2=0.88 (#186 in 1316) achi2=1.04 (#150 in 1316) [CCCCT]


Error = 1
    Quorum = 60%
      K=1 chi2=0.56 (#18487 in 35258) achi2=0.82 (#15627 in 35258) [CCCCT]
      K=2 chi2=0.00 (#34670 in 35258) achi2=0.00 (#34292 in 35258) [CCCCT]
      K=3 chi2=0.07 (#27587 in 35258) achi2=0.11 (#25832 in 35258) [CCCCT]
      K=4 chi2=0.44 (#14195 in 35258) achi2=0.44 (#14486 in 35258) [CCCCT]
```

```
5_0_msn2_4

Motivo: CCCCT
Frequência: 9/13 ~69,2%
Tamanho das Sequências: 1000

Error = 0
    Quorum = 60%
      K=1 chi2=5.34 (#66 in 826) achi2=5.23 (#72 in 826) [CCCCT]
      K=2 chi2=2.90 (#76 in 826) achi2=2.97 (#75 in 826) [CCCCT]
      K=3 chi2=1.59 (#97 in 826) achi2=1.52 (#105 in 826) [CCCCT]
      K=4 chi2=0.22 (#242 in 826) achi2=0.07 (#382 in 826) [CCCCT]
      K=5 chi2=0.00 (#360 in 826) achi2=0.00 (#339 in 826) [CCCCT]
      K=6 chi2=0.00 (#331 in 826) achi2=0.00 (#604 in 826) [CCCCT]


Error = 1
    Quorum = 60%
      K=1 chi2=3.72 (#2640 in 21544) achi2=4.62 (#2153 in 21544) [CCCCT]
      K=2 chi2=0.41 (#10937 in 21544) achi2=0.70 (#8620 in 21544) [CCCCT]
      K=3 chi2=0.00 (#20651 in 21544) achi2=0.00 (#20723 in 21544) [CCCCT]
      K=4 chi2=0.80 (#4245 in 21544) achi2=0.83 (#4322 in 21544) [CCCCT]
      K=5 chi2=0.00 (#19590 in 21562) achi2=0.00 (#20653 in 21562) [CCCCT]
      K=6 chi2=0.00 (#16728 in 21562) achi2=0.00 (#21282 in 21562) [CCCCT]


5_1_Gcr1

Motivo: CWTCC
Frequência: 10/10 ~100%
Tamanho das Sequências: 1000

Error = 0
    Quorum = 100%


Error = 1
    Quorum = 100%
      K=1 chi2=0.80 (#3213 in 6121) achi2=0.79 (#3260 in 6121) [CATCC]
      K=1 chi2=11.88 (#320 in 6121) achi2=11.58 (#344 in 6121) [CTTCC]
      K=2 chi2=1.80 (#1552 in 6121) achi2=1.61 (#1763 in 6121) [CATCC]
      K=2 chi2=5.04 (#498 in 6121) achi2=4.70 (#566 in 6121) [CTTCC]
      K=3 chi2=2.05 (#809 in 6121) achi2=1.92 (#906 in 6121) [CATCC]
      K=3 chi2=1.02 (#1713 in 6121) achi2=1.09 (#1663 in 6121) [CTTCC]
      K=4 chi2=0.97 (#929 in 6121) achi2=0.91 (#1091 in 6121) [CATCC]
      K=4 chi2=0.01 (#5090 in 6121) achi2=0.01 (#5126 in 6121) [CTTCC]
      K=5 chi2=0.00 (#5948 in 6121) achi2=0.00 (#5152 in 6121) [CATCC]
```

```
      K=5 chi2=0.00 (#4578 in 6121) achi2=0.00 (#5000 in 6121) [CTTCC]
      K=6 chi2=0.00 (#5634 in 6121) achi2=0.00 (#5438 in 6121) [CATCC]
      K=6 chi2=0.00 (#2279 in 6121) achi2=0.00 (#5149 in 6121) [CTTCC]


6_1_Gln3

Motivo: GATWAG
Frequência: 11/13 ~84,62%
Tamanho das Sequências: 1000

Error = 0
    Quorum = 80%


Error = 1
    Quorum = 80%
      K=1 chi2=2.92 (#2311 in 10819) achi2=2.73 (#2534 in 10819) [GATAAG]
      K=1 chi2=1.31 (#4204 in 10819) achi2=1.02 (#4875 in 10819) [GATTAG]
      K=2 chi2=4.99 (#766 in 10819) achi2=4.93 (#849 in 10819) [GATAAG]
      K=2 chi2=2.80 (#1760 in 10819) achi2=2.68 (#1952 in 10819) [GATTAG]
      K=3 chi2=1.68 (#2077 in 10819) achi2=1.74 (#2107 in 10819) [GATAAG]
      K=3 chi2=9.53 (#62 in 10819) achi2=9.25 (#81 in 10819) [GATTAG]
      K=4 chi2=0.14 (#6803 in 10819) achi2=0.09 (#7567 in 10819) [GATAAG]
      K=4 chi2=3.09 (#394 in 10819) achi2=3.52 (#328 in 10819) [GATTAG]
      K=5 chi2=0.01 (#8734 in 10819) achi2=0.03 (#7728 in 10819) [GATAAG]
      K=5 chi2=1.76 (#429 in 10819) achi2=1.59 (#773 in 10819) [GATTAG]
      K=6 chi2=0.00 (#9825 in 10819) achi2=0.00 (#6602 in 10819) [GATAAG]
      K=6 chi2=0.00 (#7459 in 10819) achi2=0.00 (#6760 in 10819) [GATTAG]


6_1_RAP1

Motivo: CACCCR
Frequência: 11/18 ~61,1%
Tamanho das Sequências: 1000

Error = 0
    Quorum = 60%


Error = 1
    Quorum = 60%
      K=1 chi2=3.82 (#3218 in 20531) achi2=5.55 (#2348 in 20531) [CACCCA]
      k=1 chi2=0.83 (#9754 in 20531) achi2=1.02 (#9117 in 20531) [CACCCG]
      K=2 chi2=3.30 (#2327 in 20531) achi2=3.68 (#2233 in 20531) [CACCCA]
      K=2 chi2=1.43 (#5707 in 20531) achi2=1.17 (#6956 in 20531) [CACCCG]
      K=3 chi2=2.09 (#2208 in 20531) achi2=2.23 (#2249 in 20531) [CACCCA]
```

```
      K=3 chi2=0.40 (#9516 in 20531) achi2=0.23 (#12243 in 20531) [CACCCG]
      K=4 chi2=0.71 (#4649 in 20531) achi2=1.05 (#3457 in 20531) [CACCCA]
      K=4 chi2=0.00 (#19099 in 20531) achi2=0.02 (#17523 in 20531) [CACCCG]
```

```
7_0_REB1
```

```
Motivo: TTACCC
Frequência: 7/8 ~87,5%
Tamanho das Sequências: 1000
```

```
Error = 0
    Quorum = 80%
      K=1 chi2=5.49 (#29 in 405) achi2=5.86 (#33 in 405) [TTACCC]
      K=2 chi2=6.88 (#6 in 405) achi2=6.35 (#12 in 405) [TTACCC]
      K=3 chi2=7.06 (#2 in 405) achi2=6.79 (#2 in 405) [TTACCC]
      K=4 chi2=3.73 (#3 in 405) achi2=4.02 (#3 in 405) [TTACCC]
      K=5 chi2=0.34 (#7 in 405) achi2=0.20 (#8 in 405) [TTACCC]
      K=6 chi2=0.00 (#383 in 405) achi2=0.00 (#232 in 405) [TTACCC]
```

```
Error = 1
    Quorum = 80%
      K=1 chi2=1.33 (#3134 in 10337) achi2=1.69 (#2743 in 10337) [TTACCC]
      K=2 chi2=3.02 (#657 in 10337) achi2=2.67 (#945 in 10337) [TTACCC]
      K=3 chi2=4.09 (#106 in 10337) achi2=3.94 (#170 in 10337) [TTACCC]
      K=4 chi2=0.82 (#1439 in 10337) achi2=1.09 (#1091 in 10337) [TTACCC]
      K=5 chi2=0.02 (#7204 in 10337) achi2=0.03 (#6765 in 10337) [TTACCC]
      K=6 chi2=0.00 (#10150 in 10337) achi2=0.00 (#6299 in 10337) [TTACCC]
```

```
7_0_RIM101
```

```
Motivo: TGCCAAG
Frequência: 5/7 ~71,43%
Tamanho das Sequências: 1000
```

```
Error = 0
    Quorum = 60%
      K=1 chi2=6.26 (#58 in 785) achi2=6.15 (#64 in 785) [TGCCAAG]
      K=2 chi2=6.10 (#11 in 785) achi2=5.85 (#14 in 785) [TGCCAAG]
      K=3 chi2=5.61 (#8 in 785) achi2=5.75 (#6 in 785) [TGCCAAG]
      K=4 chi2=3.99 (#4 in 785) achi2=3.77 (#6 in 785) [TGCCAAG]
      K=5 chi2=1.84 (#8 in 785) achi2=1.57 (#10 in 785) [TGCCAAG]
      K=6 chi2=0.54 (#5 in 785) achi2=0.55 (#255 in 785) [TGCCAAG]
```

```
Error = 1
    Quorum = 60%
      K=1 chi2=4.63 (#1914 in 20257) achi2=4.58 (#2166 in 20257) [TGCCAAG]
      K=2 chi2=2.92 (#1694 in 20257) achi2=2.87 (#1959 in 20257) [TGCCAAG]
      K=3 chi2=2.52 (#1158 in 20257) achi2=2.70 (#1060 in 20257) [TGCCAAG]
      K=4 chi2=0.38 (#8072 in 20257) achi2=0.30 (#9306 in 20257) [TGCCAAG]
      K=5 chi2=0.02 (#15272 in 20366) achi2=0.04 (#14013 in 20366)
[TGCCAAG]
      K=6 chi2=0.09 (#5877 in 20366) achi2=0.29 (#4589 in 20366) [TGCCAAG]


7_0_STE12

Motivo: TGAAACA
Frequência: 6/6 ~100%
Tamanho das Sequências: 1000

Error = 0
    Quorum = 100%
      K=1 chi2=10.35 (#5 in 267) achi2=9.96 (#6 in 267) [TGAAACA]
      K=2 chi2=9.61 (#1 in 267) achi2=9.05 (#1 in 267) [TGAAACA]
      K=3 chi2=9.34 (#1 in 267) achi2=8.84 (#1 in 267) [TGAAACA]
      K=4 chi2=7.94 (#1 in 267) achi2=7.31 (#1 in 267) [TGAAACA]
      K=5 chi2=2.64 (#2 in 267) achi2=3.20 (#1 in 267) [TGAAACA]
      K=6 chi2=0.44 (#3 in 267) achi2=0.52 (#168 in 267) [TGAAACA]


Error = 1
    Quorum = 100%
      K=1 chi2=4.48 (#515 in 7705) achi2=4.00 (#675 in 7705) [TGAAACA]
      K=2 chi2=1.68 (#1047 in 7705) achi2=1.57 (#1246 in 7705) [TGAAACA]
      K=3 chi2=1.59 (#637 in 7705) achi2=1.28 (#992 in 7705) [TGAAACA]
      K=4 chi2=0.57 (#1734 in 7705) achi2=0.26 (#3100 in 7705) [TGAAACA]
      K=5 chi2=0.15 (#2574 in 7705) achi2=0.21 (#2435 in 7705) [TGAAACA]
      K=6 chi2=0.07 (#1386 in 7705) achi2=0.06 (#2114 in 7705) [TGAAACA]


7_1_SWI4

Motivo: CRCGAAA
Frequência: 5/5 ~100%
Tamanho das Sequências: 1000

Error = 0
    Quorum = 100%


Error = 1
```

```
    Quorum = 100%
      K=1 chi2=9.27 (#53 in  8974) achi2=9.50 (#59 in 8974) [CACGAAA]
      K=1 chi2=14.47 (#8 in  8974) achi2=15.16 (#9 in 8974) [CGCGAAA]
      K=2 chi2=6.86 (#17 in 8974) achi2=6.38 (#22 in 8974) [CACGAAA]
      K=2 chi2=9.40 (#4 in 8974) achi2=9.24 (#3 in 8974) [CGCGAAA]
      K=3 chi2=4.56 (#46 in 8974) achi2=4.36 (#53 in 8974) [CACGAAA]
      K=3 chi2=5.14 (#20 in 8974) achi2=5.02 (#25 in 8974) [CGCGAAA]
      K=4 chi2=0.87 (#1329 in 8974) achi2=1.06 (#1095 in 8974) [CACGAAA]
      K=4 chi2=3.43 (#41 in 8974) achi2=3.90 (#27 in 8974) [CGCGAAA]
      K=5 chi2=0.12 (#3730 in 8974) achi2=0.00 (#7343 in 8974) [CACGAAA]
      K=5 chi2=0.33 (#1966 in 8974) achi2=1.21 (#397 in 8974) [CGCGAAA]
      K=6 chi2=0.07 (#1861 in 8974) achi2=0.01 (#3810 in 8974) [CACGAAA]
      K=6 chi2=0.02 (#3146 in 8974) achi2=0.44 (#873 in 8974) [CGCGAAA]

7_2_RGT1

Motivo: TGASTCW
Frequência: 0/9 ~0%
Tamanho das Sequências: 1000

Error = 0
    Quorum = 100%

Error = 1
    Quorum = 100%
      K=1 chi2=20.92 (#246 in  6085) achi2=21.40 (#236 in 6085) [CGGAAAA]
      K=1 chi2=8.10 (#799 in  6085) achi2=8.03 (#815 in 6085) [CGGAAGA]
      K=1 chi2=14.37 (#409 in 6085) achi2=14.09 (#416 in 6085) [CGGAGAA]
      K=1 chi2=3.37 (#1833 in 6085) achi2=3.50 (#1782 in 6085) [CGGATAA]
      K=2 chi2=8.37 (#251 in 6085) achi2=8.40 (#245 in 6085) [CGGAAAA]
      K=2 chi2=4.51 (#648 in 6085) achi2=4.42 (#666 in 6085) [CGGAAGA]
      K=2 chi2=9.89 (#186 in 6085) achi2=9.37 (#208 in 6085) [CGGAGAA]
      K=2 chi2=4.32 (#698 in 6085) achi2=4.28 (#695 in 6085) [CGGATAA]
      K=3 chi2=2.81 (#367 in 6085) achi2=2.72 (#385 in 6085) [CGGAAAA]
      K=3 chi2=2.42 (#473 in 6085) achi2=2.25 (#546 in 6085) [CGGAAGA]
      K=3 chi2=5.80 (#76 in 6085) achi2=5.44 (#89 in 6085) [CGGAGAA]
      K=3 chi2=3.66 (#216 in 6085) achi2=3.09 (#299 in 6085) [CGGATAA]
      K=4 chi2=0.13 (#3127 in 6085) achi2=0.22 (#2503 in 6085) [CGGAAAA]
      K=4 chi2=0.39 (#1619 in 6085) achi2=0.40 (#1722 in 6085) [CGGAAGA]
      K=4 chi2=3.20 (#35 in 6085) achi2=3.04 (#56 in 6085) [CGGAGAA]
      K=4 chi2=1.93 (#166 in 6085) achi2=2.02 (#188 in 6085) [CGGATAA]
```

```
8_1_PDR1

Motivo: CGGADDA
Frequência: 10/10 ~100%
Tamanho das Sequências: 1000

Error = 0
    Quorum = 75%
      K=1 chi2=9.01 (#25 in 639) achi2=8.84 (#27 in 639) [TCCGCGGA]
      K=2 chi2=8.86 (#13 in 639) achi2=8.83 (#14 in 639) [TCCGCGGA]
      K=3 chi2=8.62 (#4 in 639) achi2=8.64 (#5 in 639) [TCCGCGGA]
      K=4 chi2=6.59 (#2 in 639) achi2=7.41 (#2 in 639) [TCCGCGGA]
      K=5 chi2=2.52 (#1 in 639) achi2=3.06 (#1 in 639) [TCCGCGGA]
      K=6 chi2=0.12 (#5 in 639) achi2=0.30 (#253 in 639) [TCCGCGGA]


Error = 1
    Quorum = 75%
      K=1 chi2=13.88 (#184 in 16695) achi2=13.70 (#211 in 16695) [TCCGCGGA]
      K=1 chi2=9.13 (#373 in 16695) achi2=8.95 (#419 in 16695) [TCCGTGGA]
      K=2 chi2=13.55 (#68 in 16695) achi2=13.55 (#91 in 16695) [TCCGCGGA]
      K=2 chi2=9.52 (#199 in 16695) achi2=9.44 (#230 in 16695) [TCCGTGGA]
      K=3 chi2=12.37 (#5 in 16695) achi2=11.38 (#22 in 16695) [TCCGCGGA]
      K=3 chi2=7.82 (#88 in 16695) achi2=7.71 (#119 in 16695) [TCCGTGGA]
      K=4 chi2=5.70 (#45 in 16695) achi2=6.88 (#33 in 16695) [TCCGCGGA]
      K=4 chi2=5.44 (#53 in 16695) achi2=6.10 (#56 in 16695) [TCCGTGGA]
      K=5 chi2=0.61 (#2230 in 16699) achi2=1.13 (#1175 in 16699) [TCCGCGGA]
      K=5 chi2=1.01 (#981 in 16699) achi2=2.07 (#328 in 16699) [TCCGTGGA]
      K=6 chi2=0.01 (#8600 in 16699) achi2=0.00 (#10541 in 16699)
[TCCGCGGA]
      K=6 chi2=0.02 (#7065 in 16699) achi2=0.03 (#8326 in 16699) [TCCGTGGA]


8_1_PDR3

Motivo: TCCGYGGA
Frequência: 7/8 ~87,5%
Tamanho das Sequências: 1000

Error = 0
    Quorum = 80%
      K=1 chi2=9.95 (#22 in 423) achi2=9.85 (#24 in 423) [TCCGCGGA]
      K=2 chi2=9.80 (#13 in 423) achi2=9.85 (#13 in 423) [TCCGCGGA]
      K=3 chi2=9.89 (#4 in 423) achi2=9.67 (#4 in 423) [TCCGCGGA]
      K=4 chi2=7.41 (#1 in 423) achi2=8.02 (#1 in 423) [TCCGCGGA]
      K=5 chi2=2.20 (#1 in 423) achi2=3.03 (#1 in 423) [TCCGCGGA]
      K=6 chi2=0.10 (#1 in 423) achi2=0.27 (#229 in 423) [TCCGCGGA]
```

```
Error = 1
    Quorum = 80%
      K=1 chi2=16.09 (#169 in 11592) achi2=15.92 (#187 in 11592) [TCCGCGGA]
      K=1 chi2=10.90 (#327 in 11592) achi2=11.09 (#353 in 11592) [TCCGTGGA]
      K=2 chi2=16.34 (#63 in 11592) achi2=15.84 (#97 in 11592) [TCCGCGGA]
      K=2 chi2=11.48 (#174 in 11592) achi2=11.42 (#197 in 11592) [TCCGTGGA]
      K=3 chi2=13.33 (#20 in 11592) achi2=13.64 (#24 in 11592) [TCCGCGGA]
      K=3 chi2=8.78 (#91 in 11592) achi2=8.94 (#109 in 11592) [TCCGTGGA]
      K=4 chi2=7.84 (#26 in 11592) achi2=7.85 (#35 in 11592) [TCCGCGGA]
      K=4 chi2=6.32 (#50 in 11592) achi2=6.89 (#52 in 11592) [TCCGTGGA]
      K=5 chi2=0.72 (#1370 in 11601) achi2=1.27 (#778 in 11601) [TCCGCGGA]
      K=5 chi2=1.00 (#878 in 11601) achi2=1.46 (#618 in 11601) [TCCGTGGA]
      K=6 chi2=0.00 (#7969 in 11601) achi2=0.01 (#5475 in 11601) [TCCGCGGA]
      K=6 chi2=0.01 (#4626 in 11601) achi2=0.04 (#4674 in 11601) [TCCGTGGA]


8_1_PDR8

Motivo: TCCGHGGA
Frequência: 3/5 ~60%
Tamanho das Sequências: 1000

Error = 0
    Quorum = 60%
      K=1 chi2=2.85 (#128 in 1319) achi2=2.87 (#133 in 1319) [TCCGTGGA]
      K=2 chi2=2.94 (#89 in 1319) achi2=2.89 (#95 in 1319) [TCCGTGGA]
      K=3 chi2=2.83 (#10 in 1319) achi2=2.78 (#15 in 1319) [TCCGTGGA]
      K=4 chi2=2.40 (#7 in 1319) achi2=2.55 (#6 in 1319) [TCCGTGGA]
      K=5 chi2=1.22 (#16 in 1319) achi2=2.06 (#5 in 1319) [TCCGTGGA]
      K=6 chi2=0.15 (#42 in 1319) achi2=0.84 (#269 in 1319) [TCCGTGGA]

Error = 1
    Quorum = 60%
      K=1 chi2=3.36 (#1859 in 35494) achi2=3.91 (#1312 in 35494) [TCCGAGGA]
      K=1 chi2=6.12 (#489 in 35494) achi2=6.60 (#442 in 35494) [TCCGCGGA]
      K=1 chi2=3.68 (#1518 in 35494) achi2=4.02 (#1240 in 35494) [TCCGTGGA]
      K=2 chi2=3.63 (#946 in 35494) achi2=3.79 (#816 in 35494) [TCCGAGGA]
      K=2 chi2=6.36 (#218 in 35494) achi2=6.40 (#206 in 35494) [TCCGCGGA]
      K=2 chi2=4.16 (#670 in 35494) achi2=4.28 (#599 in 35494) [TCCGTGGA]
      K=3 chi2=3.01 (#498 in 35494) achi2=3.04 (#471 in 35494) [TCCGAGGA]
      K=3 chi2=5.15 (#41 in 35494) achi2=4.80 (#63 in 35494) [TCCGCGGA]
```

```
      K=3 chi2=3.41 (#278 in 35494) achi2=3.29 (#327 in 35494) [TCCGTGGA]
      K=4 chi2=2.85 (#274 in 35494) achi2=2.95 (#205 in 35494) [TCCGAGGA]
      K=4 chi2=3.00 (#198 in 35494) achi2=3.37 (#85 in 35494) [TCCGCGGA]
      K=4 chi2=2.17 (#1082 in 35494) achi2=2.19 (#1060 in 35494) [TCCGTGGA]
      K=5 chi2=0.75 (#6352 in 35725) achi2=0.87 (#6044 in 35725) [TCCGAGGA]
      K=5 chi2=0.79 (#5926 in 35725) achi2=1.53 (#2215 in 35725) [TCCGCGGA]
      K=5 chi2=0.81 (#5763 in 35725) achi2=0.80 (#6754 in 35725) [TCCGTGGA]
      K=6 chi2=0.03 (#19985 in 35725) achi2=0.10 (#17420 in 35725)
[TCCGAGGA]
      K=6 chi2=0.26 (#6383 in 35725) achi2=0.53 (#6615 in 35725) [TCCGCGGA]
      K=6 chi2=0.11 (#11922 in 35725) achi2=0.00 (#30343 in 35725)
[TCCGTGGA]
```

Anexo C Resultados do Cálculo da Significância Estatística referente à Abundância pelo Método Analítico para Conjuntos de Dados de *Saccharomyces cerevisiae*

Foi feita uma procura, usando o SMILE, em vários conjuntos de dados.

A parametrização usada fez variar os seguintes parâmetros:

k = 1,...,4
erro = 0,1
quórum = 60, 75, 80, 100

A procura foi feita para motivos com comprimento de 4 a 10, excepto no conjunto de dados ZAP1 onde a procura foi feita para tamanhos entre 4 e 11.

De seguida apresentamos os resultados para os motivos procurados em relação a cada valor dos parâmetros.

Descrição do formato dos resultados:

K=<*k-mer* conservado> achi2 = <chi2 analítico> (<posição> in <número de motivos extraídos> {<posição entre motivos do mesmo tamanho>} [motivo]

Nota: A posição do motivo refere-se à ordem em que se encontra numa lista ordenada por ordem decrescente de chi2 analítico dos motivos extraídos.

```
10_2_YRR1

Motivo: TTCCGTGGAA   TTCCGCGGAT    TTCCGCGGAA
Tamanho das Sequências: 1000

Error = 0
    Quorum = 100%

Error = 1
    Quorum = 100%

      K=1 achi2=4.76 (#745 in 8908) {#13} [TTCCGCGGAA]
      K=2 achi2=4.64 (#509 in 8908) {#12} [TTCCGCGGAA]
      K=3 achi2=4.13 (#230 in 8908) {#11} [TTCCGCGGAA]
      K=4 achi2=2.78 (#205 in 8908) {#10} [TTCCGCGGAA]
      K=5 achi2=1.43 (#287 in 8908) {#12} [TTCCGCGGAA]
      K=6 achi2=0.41 (#1031 in 8908) {#15} [TTCCGCGGAA]
      K=7 achi2=0.00 (#2230 in 8908) {#16} [TTCCGCGGAA]
      K=8 achi2=0.01 (#4728 in 8908) {#16} [TTCCGCGGAA]
      K=9 achi2=0.01 (#7926 in 8908) {#16} [TTCCGCGGAA]
      K=10 achi2=0.00 (#8782 in 8908) {#16} [TTCCGCGGAA]
```

```
      K=11 achi2=0.00 (#8903 in 8908) {#12} [TTCCGCGGAA]

11_2_ZAP1

Motivo: ACCTTGAAGGT     ACCCTAAAGGT     ACCTTAAAGGT
Frequência: 4/5 ~80%
Tamanho das Sequências: 1000

Error = 0
    Quorum = 80%

Error = 1
    Quorum = 80%

      K=1 achi2=3.87 (#504 in 17722) {#3} [ACCCTAAAGGT]
      K=1 achi2=4.83 (#265 in 17722) {#2} [ACCTTAAAGGT]
      K=2 achi2=3.89 (#248 in 17722) {#3} [ACCCTAAAGGT]
      K=2 achi2=4.84 (#119 in 17722) {#2} [ACCTTAAAGGT]
      K=3 achi2=3.91 (#61 in 17722) {#3} [ACCCTAAAGGT]
      K=3 achi2=4.85 (#14 in 17722) {#2} [ACCTTAAAGGT]
      K=4 achi2=3.87 (#31 in 17722) {#4} [ACCCTAAAGGT]
      K=4 achi2=4.85 (#4 in 17722) {#2} [ACCTTAAAGGT]
      K=5 achi2=3.43 (#29 in 17722) {#4} [ACCCTAAAGGT]
      K=5 achi2=4.62 (#3 in 17722) {#2} [ACCTTAAAGGT]
      K=6 achi2=2.82 (#276 in 17722) {#4} [ACCCTAAAGGT]
      K=6 achi2=4.31 (#259 in 17722) {#2} [ACCTTAAAGGT]
      K=7 achi2=0.63 (#1455 in 17722) {#4} [ACCCTAAAGGT]
      K=7 achi2=1.97 (#1286 in 17722) {#3} [ACCTTAAAGGT]
      K=8 achi2=0.04 (#5422 in 17722) {#3} [ACCCTAAAGGT]
      K=8 achi2=0.05 (#5392 in 17722) {#1} [ACCTTAAAGGT]
      K=9 achi2=0.00 (#13376 in 17722) {#3} [ACCCTAAAGGT]
      K=9 achi2=0.05 (#13364 in 17722) {#1} [ACCTTAAAGGT]
      K=10 achi2=0.00 (#17186 in 17722) {#3} [ACCCTAAAGGT]
      K=10 achi2=0.00 (#17184 in 17722) {#1} [ACCTTAAAGGT]
      K=11 achi2=0.00 (#17688 in 17722) {#3} [ACCCTAAAGGT]
      K=11 achi2=0.00 (#17686 in 17722) {#1} [ACCTTAAAGGT]
```

```
5_0_msn2_4

Motivo: CCCCT
Frequência: 9/13 ~69,2%
Tamanho das Sequências: 1000

Error = 0
    Quorum = 60%
      K=1 achi2=5.23 (#72 in 826) {#21} [CCCCT]
      K=2 achi2=2.97 (#75 in 826) {#23} [CCCCT]
      K=3 achi2=1.52 (#105 in 826) {#55} [CCCCT]
      K=4 achi2=0.07 (#382 in 826) {#310} [CCCCT]
      K=5 achi2=0.00 (#339 in 826) {#245} [CCCCT]
      K=6 achi2=0.00 (#604 in 826) {#255} [CCCCT]


Error = 1
    Quorum = 60%
      K=1 achi2=4.62 (#2153 in 21544) {#223} [CCCCT]
      K=2 achi2=0.70 (#8620 in 21544) {#476} [CCCCT]
      K=3 achi2=0.00 (#20723 in 21544) {#986} [CCCCT]
      K=4 achi2=0.83 (#4322 in 21544) {#81} [CCCCT]
      K=5 achi2=0.00 (#20635 in 21544) {#618} [CCCCT]
      K=6 achi2=0.00 (#21264 in 21544) {#744} [CCCCT]



5_0_MSN4

Motivo: CCCCT
Frequência: 3/5 ~60%
Tamanho das Sequências: 1000

Error = 0
    Quorum = 60%
      K=1 achi2=3.14 (#116 in 1316) {#27} [CCCCT]
      K=2 achi2=1.94 (#178 in 1316) {#39} [CCCCT]
      K=3 achi2=1.40 (#190 in 1316) {#42} [CCCCT]
      K=4 achi2=1.04 (#150 in 1316) {#19} [CCCCT]
      K=5 achi2=0.00 (#576 in 1316) {#111} [CCCCT]
      K=6 achi2=0.00 (#1113 in 1316) {#392} [CCCCT]
      K=7 achi2=9.01 (#324 in 1316) {#94} [CCCCT]
      K=8 achi2=9.01 (#340 in 1316) {#94} [CCCCT]
      K=9 achi2=9.01 (#341 in 1316) {#94} [CCCCT]
      K=10 achi2=9.01 (#342 in 1316) {#94} [CCCCT]
      K=11 achi2=9.01 (#343 in 1316) {#94} [CCCCT]
```

```
Error = 1
    Quorum = 60%
      K=1 achi2=0.82 (#15627 in 35258) {#568} [CCCCT]
      K=2 achi2=0.00 (#34292 in 35258) {#986} [CCCCT]
      K=3 achi2=0.11 (#25832 in 35258) {#671} [CCCCT]
      K=4 achi2=0.44 (#14486 in 35258) {#144} [CCCCT]
      K=5 achi2=0.00 (#34387 in 35258) {#626} [CCCCT]
      K=6 achi2=0.00 (#34560 in 35258) {#326} [CCCCT]


5_1_Gcr1

Motivo: CWTCC
Frequência: 10/10 ~100%
Tamanho das Sequências: 1000

Error = 0
    Quorum = 100%
Error = 1
    Quorum = 100%

      K=1 achi2=0.79 (#3260 in 6121) {#599} [CATCC]
      K=1 achi2=11.58 (#344 in 6121) {#50} [CTTCC]
      K=2 achi2=1.61 (#1763 in 6121) {#312} [CATCC]
      K=2 achi2=4.70 (#566 in 6121) {#105} [CTTCC]
      K=3 achi2=1.92 (#906 in 6121) {#128} [CATCC]
      K=3 achi2=1.09 (#1663 in 6121) {#248} [CTTCC]
      K=4 achi2=0.91 (#1091 in 6121) {#97} [CATCC]
      K=4 achi2=0.01 (#5126 in 6121) {#870} [CTTCC]
      K=5 achi2=0.00 (#5152 in 6121) {#372} [CATCC]
      K=5 achi2=0.00 (#5000 in 6121) {#230} [CTTCC]
      K=6 achi2=0.00 (#5438 in 6121) {#341} [CATCC]
      K=6 achi2=0.00 (#5149 in 6121) {#52} [CTTCC]
```

```
6_1_Gln3

Motivo: GATWAG
Frequência: 11/13 ~84,62%
Tamanho das Sequências: 1000


Error = 0
    Quorum = 80%
Error = 1
    Quorum = 80%

      K=1 achi2=2.73 (#2534 in 10819) {#831} [GATAAG]
      K=1 achi2=1.02 (#4875 in 10819) {#1698} [GATTAG]
      K=2 achi2=4.93 (#849 in 10819) {#251} [GATAAG]
      K=2 achi2=2.68 (#1952 in 10819) {#593} [GATTAG]
      K=3 achi2=1.74 (#2107 in 10819) {#675} [GATAAG]
      K=3 achi2=9.25 (#81 in 10819) {#12} [GATTAG]
      K=4 achi2=0.09 (#7567 in 10819) {#2712} [GATAAG]
      K=4 achi2=3.52 (#328 in 10819) {#59} [GATTAG]
      K=5 achi2=0.03 (#7728 in 10819) {#3046} [GATAAG]
      K=5 achi2=1.59 (#773 in 10819) {#110} [GATTAG]
      K=6 achi2=0.00 (#6602 in 10819) {#761} [GATAAG]
      K=6 achi2=0.00 (#6760 in 10819) {#917} [GATTAG]
      K=7 achi2=0.00 (#10606 in 10819) {#3684} [GATAAG]
      K=7 achi2=0.00 (#8570 in 10819) {#1648} [GATTAG]



6_1_RAP1

Motivo: CACCCR
Frequência: 11/18 ~61,1%
Tamanho das Sequências: 1000

Error = 0
    Quorum = 60%


Error = 1
    Quorum = 60%
      K=1 achi2=5.55 (#2348 in 20531) {#456} [CACCCA]
      K=1 achi2=1.02 (#9117 in 20531) {#2005} [CACCCG]
      K=2 achi2=3.68 (#2233 in 20531) {#460} [CACCCA]
      K=2 achi2=1.17 (#6956 in 20531) {#1498} [CACCCG]
      K=3 achi2=2.23 (#2249 in 20531) {#392} [CACCCA]
```

```
      K=3 achi2=0.23 (#12243 in 20531) {#2397} [CACCCG]
      K=4 achi2=1.05 (#3457 in 20531) {#482} [CACCCA]
      K=4 achi2=0.02 (#17523 in 20531) {#3457} [CACCCG]
      K=5 achi2=0.06 (#13001 in 20531) {#2409} [CACCCA]
      K=5 achi2=0.48 (#5016 in 20531) {#610} [CACCCG]
      K=6 achi2=0.00 (#17016 in 20531) {#1751} [CACCCA]
      K=6 achi2=0.00 (#18636 in 20531) {#3343} [CACCCG]
      K=7 achi2=0.00 (#17654 in 20531) {#1218} [CACCCA]
      K=7 achi2=0.00 (#18327 in 20531) {#1891} [CACCCG]

7_0_REB1

Motivo: TTACCC
Frequência: 7/8 ~87,5%
Tamanho das Sequências: 1000

Error = 0
    Quorum = 80%
      K=1 achi2=5.86 (#33 in 405) {#5} [TTACCC]
      K=2 achi2=6.35 (#12 in 405) {#1} [TTACCC]
      K=3 achi2=6.79 (#2 in 405) {#1} [TTACCC]
      K=4 achi2=4.02 (#3 in 405) {#2} [TTACCC]
      K=5 achi2=0.20 (#8 in 405) {#7} [TTACCC]
      K=6 achi2=0.00 (#232 in 405) {#8} [TTACCC]
      K=7 achi2=0.00 (#402 in 405) {#8} [TTACCC]

Error = 1
    Quorum = 80%
      K=1 achi2=1.69 (#2742 in 10337) {#933} [TTACCC]
      K=2 achi2=2.67 (#945 in 10337) {#284} [TTACCC]
      K=3 achi2=3.94 (#170 in 10337) {#25} [TTACCC]
      K=4 achi2=1.09 (#1091 in 10337) {#246} [TTACCC]
      K=5 achi2=0.03 (#6765 in 10337) {#2619} [TTACCC]
      K=6 achi2=0.00 (#6295 in 10337) {#852} [TTACCC]
      K=7 achi2=0.00 (#9863 in 10337) {#3333} [TTACCC]
```

```
7_0_RIM101


Motivo: TGCCAAG
Frequência: 5/7 ~71,43%
Tamanho das Sequências: 1000

Error = 0
    Quorum = 60%

      K=1 achi2=6.15 (#64 in 785) {#12} [TGCCAAG]
      K=2 achi2=5.85 (#14 in 785) {#3} [TGCCAAG]
      K=3 achi2=5.75 (#6 in 785) {#2} [TGCCAAG]
      K=4 achi2=3.77 (#6 in 785) {#2} [TGCCAAG]
      K=5 achi2=1.57 (#10 in 785) {#5} [TGCCAAG]
      K=6 achi2=0.55 (#255 in 785) {#4} [TGCCAAG]
      K=7 achi2=0.00 (#652 in 785) {#18} [TGCCAAG]
      K=8 achi2=0.00 (#771 in 785) {#15} [TGCCAAG]

Error = 1
    Quorum = 60%

      K=1 achi2=4.58 (#2166 in 20257) {#407} [TGCCAAG]
      K=2 achi2=2.87 (#1959 in 20257) {#462} [TGCCAAG]
      K=3 achi2=2.70 (#1060 in 20257) {#241} [TGCCAAG]
      K=4 achi2=0.30 (#9306 in 20257) {#3695} [TGCCAAG]
      K=5 achi2=0.04 (#13915 in 20257) {#6353} [TGCCAAG]
      K=6 achi2=0.29 (#4512 in 20257) {#1713} [TGCCAAG]
      K=7 achi2=0.03 (#5275 in 20257) {#36} [TGCCAAG]
      K=8 achi2=0.00 (#12401 in 20257) {#1099} [TGCCAAG]
7_0_STE12
Motivo: TGAAACA
Frequência: 6/6 ~100%
Tamanho das Sequências: 1000

Error = 0
    Quorum = 100%

      K=1 achi2=9.96 (#6 in 267) {#1} [TGAAACA]
      K=2 achi2=9.05 (#1 in 267) {#1} [TGAAACA]
      K=3 achi2=8.84 (#1 in 267) {#1} [TGAAACA]
      K=4 achi2=7.31 (#1 in 267) {#1} [TGAAACA]
      K=5 achi2=3.20 (#1 in 267) {#1} [TGAAACA]
```

```
      K=6 achi2=0.52 (#168 in 267) {#2} [TGAAACA]
      K=7 achi2=0.00 (#254 in 267) {#1} [TGAAACA]
      K=8 achi2=0.00 (#266 in 267) {#2} [TGAAACA]

Error = 1
    Quorum = 100%

      K=1 achi2=4.00 (#675 in 7705) {#162} [TGAAACA]
      K=2 achi2=1.57 (#1246 in 7705) {#402} [TGAAACA]
      K=3 achi2=1.28 (#992 in 7705) {#375} [TGAAACA]
      K=4 achi2=0.26 (#3100 in 7705) {#1267} [TGAAACA]
      K=5 achi2=0.21 (#2435 in 7705) {#1154} [TGAAACA]
      K=6 achi2=0.06 (#2114 in 7705) {#1435} [TGAAACA]
      K=7 achi2=0.00 (#2095 in 7705) {#260} [TGAAACA]
      K=8 achi2=0.00 (#6802 in 7705) {#1861} [TGAAACA]

7_1_SWI4

Motivo: CRCGAAA
Frequência: 5/5 ~100%
Tamanho das Sequências: 1000

Error = 0
    Quorum = 100%

Error = 1
    Quorum = 100%
      K=1 achi2=9.50 (#59 in 8974) {#15} [CACGAAA]
      K=1 achi2=15.16 (#9 in 8974) {#2} [CGCGAAA]
      K=2 achi2=6.38 (#22 in 8974) {#7} [CACGAAA]
      K=2 achi2=9.24 (#3 in 8974) {#1} [CGCGAAA]
      K=3 achi2=4.36 (#53 in 8974) {#25} [CACGAAA]
      K=3 achi2=5.02 (#25 in 8974) {#10} [CGCGAAA]
      K=4 achi2=1.06 (#1095 in 8974) {#476} [CACGAAA]
      K=4 achi2=3.90 (#27 in 8974) {#9} [CGCGAAA]
      K=5 achi2=0.00 (#7343 in 8974) {#3297} [CACGAAA]
      K=5 achi2=1.21 (#397 in 8974) {#176} [CGCGAAA]
      K=6 achi2=0.01 (#3810 in 8974) {#2702} [CACGAAA]
      K=6 achi2=0.44 (#873 in 8974) {#314} [CGCGAAA]
      K=7 achi2=0.00 (#2384 in 8974) {#205} [CACGAAA]
      K=7 achi2=0.00 (#2371 in 8974) {#194} [CGCGAAA]
      K=8 achi2=0.00 (#8531 in 8974) {#3025} [CACGAAA]
```

```
      K=8 achi2=0.00 (#5630 in 8974) {#124} [CGCGAAA]

7_2_RGT1

Motivo: TGASTCW
Frequência: 0/9 ~0%
Tamanho das Sequências: 1000

Error = 0
    Quorum = 100%

Error = 1
    Quorum = 100%

      K=1 achi2=21.40 (#236 in 6085) {#74} [CGGAAAA]
      K=1 achi2=8.03 (#815 in 6085) {#253} [CGGAAGA]
      K=1 achi2=14.09 (#416 in 6085) {#129} [CGGAGAA]
      K=1 achi2=3.50 (#1782 in 6085) {#511} [CGGATAA]
      K=2 achi2=8.40 (#245 in 6085) {#67} [CGGAAAA]
      K=2 achi2=4.42 (#666 in 6085) {#169} [CGGAAGA]
      K=2 achi2=9.37 (#208 in 6085) {#55} [CGGAGAA]
      K=2 achi2=4.28 (#695 in 6085) {#175} [CGGATAA]
      K=3 achi2=2.72 (#385 in 6085) {#126} [CGGAAAA]
      K=3 achi2=2.25 (#546 in 6085) {#173} [CGGAAGA]
      K=3 achi2=5.44 (#89 in 6085) {#28} [CGGAGAA]
      K=3 achi2=3.09 (#299 in 6085) {#100} [CGGATAA]
      K=4 achi2=0.22 (#2503 in 6085) {#821} [CGGAAAA]
      K=4 achi2=0.40 (#1722 in 6085) {#598} [CGGAAGA]
      K=4 achi2=3.04 (#56 in 6085) {#23} [CGGAGAA]
      K=4 achi2=2.02 (#188 in 6085) {#81} [CGGATAA]
      K=5 achi2=0.03 (#3495 in 6085) {#1206} [CGGAAAA]
      K=5 achi2=0.19 (#1882 in 6085) {#749} [CGGAAGA]
      K=5 achi2=2.38 (#34 in 6085) {#12} [CGGAGAA]
      K=5 achi2=1.38 (#154 in 6085) {#74} [CGGATAA]
      K=6 achi2=0.09 (#1192 in 6085) {#744} [CGGAAAA]
      K=6 achi2=0.31 (#682 in 6085) {#300} [CGGAAGA]
      K=6 achi2=0.57 (#459 in 6085) {#122} [CGGAGAA]
      K=6 achi2=0.42 (#550 in 6085) {#191} [CGGATAA]
      K=7 achi2=0.00 (#1608 in 6085) {#53} [CGGAAAA]
      K=7 achi2=0.00 (#2439 in 6085) {#861} [CGGAAGA]
      K=7 achi2=0.00 (#2150 in 6085) {#572} [CGGAGAA]
      K=7 achi2=0.00 (#2472 in 6085) {#894} [CGGATAA]
```

```
      K=8 achi2=0.00 (#6056 in 6085) {#1512} [CGGAAAA]
      K=8 achi2=0.00 (#5728 in 6085) {#1184} [CGGAAGA]
      K=8 achi2=0.00 (#5127 in 6085) {#583} [CGGAGAA]
      K=8 achi2=0.00 (#5961 in 6085) {#1417} [CGGATAA]

8_1_PDR1

Motivo: CGGADDA
Frequência: 10/10 ~100%
Tamanho das Sequências: 1000

Error = 0
    Quorum = 75%
      K=1 achi2=8.84 (#27 in 639) {#1} [TCCGCGGA]
      K=2 achi2=8.83 (#14 in 639) {#1} [TCCGCGGA]
      K=3 achi2=8.64 (#5 in 639) {#1} [TCCGCGGA]
      K=4 achi2=7.41 (#2 in 639) {#1} [TCCGCGGA]
      K=5 achi2=3.06 (#1 in 639) {#1} [TCCGCGGA]
      K=6 achi2=0.30 (#253 in 639) {#1} [TCCGCGGA]
      K=7 achi2=0.12 (#587 in 639) {#1} [TCCGCGGA]
      K=8 achi2=0.00 (#633 in 639) {#1} [TCCGCGGA]
      K=9 achi2=0.00 (#639 in 639) {#1} [TCCGCGGA]

Error = 1
    Quorum = 75%

      K=1 achi2=13.70 (#211 in 16695) {#48} [TCCGCGGA]
      K=1 achi2=8.95 (#419 in 16695) {#111} [TCCGTGGA]
      K=2 achi2=13.55 (#91 in 16695) {#26} [TCCGCGGA]
      K=2 achi2=9.44 (#230 in 16695) {#59} [TCCGTGGA]
      K=3 achi2=11.38 (#22 in 16695) {#13} [TCCGCGGA]
      K=3 achi2=7.71 (#119 in 16695) {#39} [TCCGTGGA]
      K=4 achi2=6.88 (#33 in 16695) {#9} [TCCGCGGA]
      K=4 achi2=6.10 (#56 in 16695) {#18} [TCCGTGGA]
      K=5 achi2=1.13 (#1171 in 16695) {#496} [TCCGCGGA]
      K=5 achi2=2.07 (#324 in 16695) {#135} [TCCGTGGA]
      K=6 achi2=0.00 (#10537 in 16695) {#2237} [TCCGCGGA]
      K=6 achi2=0.03 (#8322 in 16695) {#1955} [TCCGTGGA]
      K=7 achi2=0.00 (#4060 in 16695) {#2312} [TCCGCGGA]
      K=7 achi2=0.00 (#4066 in 16695) {#2318} [TCCGTGGA]
      K=8 achi2=0.00 (#5835 in 16695) {#174} [TCCGCGGA]
      K=8 achi2=0.00 (#6050 in 16695) {#389} [TCCGTGGA]
```

```
      K=9 achi2=0.00 (#15679 in 16695) {#1375} [TCCGCGGA]
      K=9 achi2=0.00 (#14892 in 16695) {#588} [TCCGTGGA]

8_1_PDR3

Motivo: TCCGYGGA
Frequência: 7/8 ~87,5%
Tamanho das Sequências: 1000

Error = 0
    Quorum = 80%
      K=1 achi2=9.85 (#24 in 423) {#1} [TCCGCGGA]
      K=2 achi2=9.85 (#13 in 423) {#1} [TCCGCGGA]
      K=3 achi2=9.67 (#4 in 423) {#1} [TCCGCGGA]
      K=4 achi2=8.02 (#1 in 423) {#1} [TCCGCGGA]
      K=5 achi2=3.03 (#1 in 423) {#1} [TCCGCGGA]
      K=6 achi2=0.27 (#229 in 423) {#1} [TCCGCGGA]
      K=7 achi2=0.13 (#401 in 423) {#1} [TCCGCGGA]
      K=8 achi2=0.00 (#420 in 423) {#1} [TCCGCGGA]
      K=9 achi2=0.00 (#423 in 423) {#1} [TCCGCGGA]

Error = 1
    Quorum = 80%

      K=1 achi2=15.92 (#187 in 11592) {#35} [TCCGCGGA]
      K=1 achi2=11.09 (#353 in 11592) {#83} [TCCGTGGA]
      K=2 achi2=15.84 (#97 in 11592) {#25} [TCCGCGGA]
      K=2 achi2=11.42 (#197 in 11592) {#44} [TCCGTGGA]
      K=3 achi2=13.64 (#24 in 11592) {#9} [TCCGCGGA]
      K=3 achi2=8.94 (#109 in 11592) {#35} [TCCGTGGA]
      K=4 achi2=7.85 (#35 in 11592) {#11} [TCCGCGGA]
      K=4 achi2=6.89 (#52 in 11592) {#16} [TCCGTGGA]
      K=5 achi2=1.27 (#769 in 11592) {#243} [TCCGCGGA]
      K=5 achi2=1.46 (#609 in 11592) {#204} [TCCGTGGA]
      K=6 achi2=0.01 (#5466 in 11592) {#879} [TCCGCGGA]
      K=6 achi2=0.04 (#4665 in 11592) {#789} [TCCGTGGA]
      K=7 achi2=0.00 (#2570 in 11592) {#968} [TCCGCGGA]
      K=7 achi2=0.00 (#2517 in 11592) {#941} [TCCGTGGA]
      K=8 achi2=0.00 (#5547 in 11592) {#157} [TCCGCGGA]
      K=8 achi2=0.00 (#5666 in 11592) {#276} [TCCGTGGA]
      K=9 achi2=0.00 (#10744 in 11592) {#135} [TCCGCGGA]
      K=9 achi2=0.00 (#11041 in 11592) {#432} [TCCGTGGA]
```

```
8_1_PDR8

Motivo: TCCGHGGA
Frequência: 3/5 ~60%
Tamanho das Sequências: 1000


Error = 0
    Quorum = 60%

      K=1 achi2=2.87 (#133 in 1319) {#8} [TCCGTGGA]
      K=2 achi2=2.89 (#95 in 1319) {#5} [TCCGTGGA]
      K=3 achi2=2.78 (#15 in 1319) {#2} [TCCGTGGA]
      K=4 achi2=2.55 (#6 in 1319) {#1} [TCCGTGGA]
      K=5 achi2=2.06 (#5 in 1319) {#2} [TCCGTGGA]
      K=6 achi2=0.84 (#269 in 1319) {#7} [TCCGTGGA]
      K=7 achi2=0.00 (#897 in 1319) {#17} [TCCGTGGA]
      K=8 achi2=0.00 (#1232 in 1319) {#8} [TCCGTGGA]
      K=9 achi2=0.00 (#1303 in 1319) {#1} [TCCGTGGA]

Error = 1
    Quorum = 60%

      K=1 achi2=3.91 (#1312 in 35494) {#312} [TCCGAGGA]
      K=1 achi2=6.60 (#442 in 35494) {#89} [TCCGCGGA]
      K=1 achi2=4.02 (#1240 in 35494) {#301} [TCCGTGGA]
      K=2 achi2=3.79 (#816 in 35494) {#174} [TCCGAGGA]
      K=2 achi2=6.40 (#206 in 35494) {#48} [TCCGCGGA]
      K=2 achi2=4.28 (#599 in 35494) {#123} [TCCGTGGA]
      K=3 achi2=3.04 (#471 in 35494) {#120} [TCCGAGGA]
      K=3 achi2=4.80 (#63 in 35494) {#10} [TCCGCGGA]
      K=3 achi2=3.29 (#327 in 35494) {#87} [TCCGTGGA]
      K=4 achi2=2.95 (#205 in 35494) {#61} [TCCGAGGA]
      K=4 achi2=3.37 (#85 in 35494) {#28} [TCCGCGGA]
      K=4 achi2=2.19 (#1060 in 35494) {#257} [TCCGTGGA]
      K=5 achi2=0.87 (#5828 in 35494) {#2082} [TCCGAGGA]
      K=5 achi2=1.53 (#2062 in 35494) {#521} [TCCGCGGA]
      K=5 achi2=0.80 (#6534 in 35494) {#2399} [TCCGTGGA]
      K=6 achi2=0.10 (#17189 in 35494) {#7240} [TCCGAGGA]
      K=6 achi2=0.53 (#6392 in 35494) {#2378} [TCCGCGGA]
      K=6 achi2=0.00 (#30112 in 35494) {#12027} [TCCGTGGA]
      K=7 achi2=0.06 (#9736 in 35494) {#4698} [TCCGAGGA]
      K=7 achi2=0.06 (#10113 in 35494) {#5025} [TCCGCGGA]
```

```
K=7 achi2=0.00 (#17876 in 35494) {#11124} [TCCGTGGA]
K=8 achi2=0.00 (#12067 in 35494) {#1424} [TCCGAGGA]
K=8 achi2=0.00 (#11172 in 35494) {#529} [TCCGCGGA]
K=8 achi2=0.00 (#11688 in 35494) {#1045} [TCCGTGGA]
K=9 achi2=0.00 (#31497 in 35494) {#8200} [TCCGAGGA]
K=9 achi2=0.00 (#25674 in 35494) {#2377} [TCCGCGGA]
K=9 achi2=0.00 (#29818 in 35494) {#6521} [TCCGTGGA]
```