Universidade Federal Fluminense
Instituto de Computação
Coordenação do Curso de Pós-Graduação em Computação

RESUMO DA APRESENTAÇÃO E RELATÓRIO DA IMPLEMENTAÇÃO

TEMA: Identificação de sinais biológicos relevantes em sequências de DNA

Aluno: Jivago Medeiros (*jmedeiros@ic.uff.br*)

Professora: Helena Cristina

MOTIVAÇÃO

A principal motivação para a escolha do tema foi por pessoalmente, dentre os conteúdos apresentados durante a disciplina de Introdução a Biologia Molecular Computacional, o conteúdo relacionado a predição de genes *(gene prediction)* ter despertado maior interesse. Acredito que a predição de gênica ainda hoje é um problema de bastante relevância para a computação e que realizar um estudo que relacionasse o processamento de sinais digitais ao tema seria de bastante valia para mim e também para os colegas da disciplina que por ventura assistissem minha apresentação.

RESUMO DA APRESENTAÇÃO

Este resumo, refere-se a apresentação que realizei em sala de aula do trabalho da Tiwari *et al.* 1997 e tendo também o trabalho da Pessoa 2004 como um trabalho que forneceu suporte ao entendimento de alguns conceitos relacionados a utilização de processamento de sinais digitais (principalmente Análise de Fourier) em biologia molecular.

A predição de regiões codificadoras em sequências de DNA é ainda hoje um dos principais problemas computacionais na biologia molecular. Em 1997, Tiwari *et al.* apontavam o fato da maioria das estratégias utilizadas para esse fim necessitarem das características de regiões codificadoras e não codificadoras já conhecidas. Como por exemplo padrões de códons, frequência de oligonucleotídios ou mesmo redes neurais. As redes neurais por exemplo necessitam que seja realizado treinamento com regiões codificadoras e não codificadores de sequências já conhecidas, e após o treinamento podem produzir resultados satisfatórios desde de que as sequências testadas sejam da mesma espécie.

Fickett 1982 apresentou um estudo onde são listadas diversas características de regiões codificadoras e não codificadoras. Uma dessas característica forneceu a base necessária para a utilização de processamento de sinais digitais como ferramenta na predição de genes; a periodicidade três das regiões codificadoras. Nas regiões codificadoras, uma determinada base se repete com período 3 (ou frequência = 1/3), característica essa que não está presente em regiões codificadoras. A presença de periodicidade 3 em regiões codificadoras e a ausência em regiões não codificadoras é considerada uma característica universal das sequências de DNA (Tiware *et al.* 1997).

Voss 1992 demonstrou que era possível observar a característica da periodicidade 3 utilizando Análise de Fourier em uma sequência de DNA. A característica estaria visível por meio de um pico espectral na frequência 1/3. A Figura 1 demonstra o pico no espectro na frequência 1/3 para uma região codificadora em (a) e em (b) a ausência desse pico para uma região não codificadora.

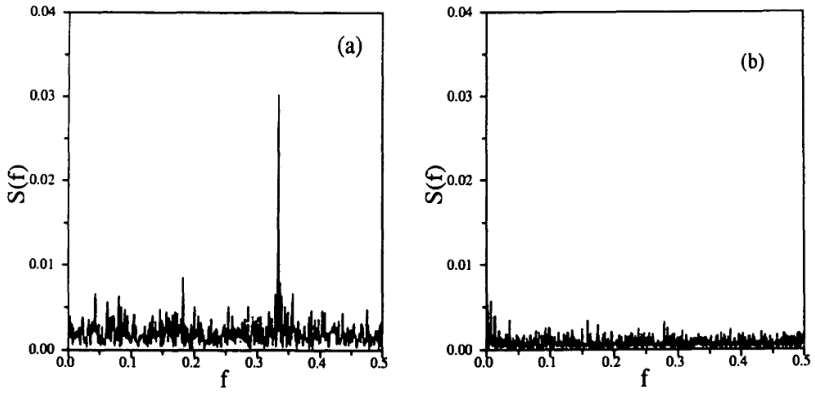

Figura 1. Presença em (a) do pico em *f*=1/3 no espectro e ausência em (b). **Fonte:** Tiwari *et al.* 1997

A transformada de Fourier é um método matemático proposto por Joseh Fourier que altera uma função do seu domínio amostral (por exmplo o tempo) para o domínio da frequência. Após plotar um gráfico de uma sequência obtida pela Transformada de Fourier, temos no eixo *x* as frequências e no eixo *y* a amplitude dessas frequências. A Figura 2 ilustra a utilização da Transformada de Fourier para alterar o domínio de uma amostra para o domínio da frequência e a Transformada Inversa de Fourier para voltar para o domínio amostral inicial.

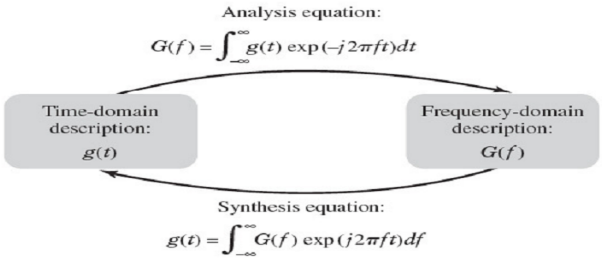

Figura 2. Utilização da Transforma de Fourier para alterar o domínio de uma amostra **Fonte:** desconhecida

A fórmula para aplicação da Transformada de Fourier e Transformada Inversa de Fourier está indicada na Figura 2. Porém, em processamento de sinais digitais trabalhos com intervalos discretos, e não contínuos (Transformada de Fourier). Assim, no processamento de sinais digitais utilizamos a Transformada Discreta de Fourier (TDF), que utiliza somatório ao invés de integral e cuja formulá está indicada na Figura 3.

$$v(m) = \frac{1}{N} \sum_{k=0}^{N-1} u(k) \exp[-j2\pi mk/N]$$

Figura 3. Transformada Discreta de Fourier **Fonte:** Wikipedia

Uma das principais vantagens da utilização da Transforma de Fourier é que ela consegue trabalhar de maneira eficiente com as correlações existentes entre as bases de uma sequência de DNA. O que não acontece com muitos outros métodos matemáticos e computacionais. Todos os coeficientes de Fourier em uma sequência são variáveis aleatórias independentes, o que faz com que a correlação entre bases próximas não seja um obstáculo quando se trabalho com a Transformada de Fourier (Pessoa, 2004).

Como mencionado anteriormente, a transformada de Fourier aplicada a uma sequência de base nitrogenadas é capaz de indicar a periodicidade 3 das regiões codificadoras do DNA. Porém, para processar uma sequência de bases, que tradicionalmente são representadas pelas letras AGCT, é necessário que as bases de uma sequência sejam representadas por valores numéricos e contínuos. Há documentado algumas maneira de realizar essa representação. Talvez a mais intuitiva e simples seja substituir as bases AGCT por 1234 (ou ainda 0123). Porém, dessa forma estaríamos dizendo que determinada base tem "peso" maior sobre outra na amostra.

Um método muito utilizado até hoje é o proposto por em Voss 1992 (Figura 4) que gera pra cada base – AGCT – um vetor *Ua*, *Ug*, *Uc*, *Ut* respectivamente, onde o tamanho de cada vetor é o tamanho total da sequência e em cada posição de cada vetor, utiliza-se 1 para dizer que naquela posição da sequência é a base corresponde ao vetor que está presente conforme pode ser visto na figura a seguir.

Sequence	G	G	A	T	A	T	C	A	C	T	T	T	A	G	A	G
Apply U_A	0	0	1	0	1	0	0	1	0	0	0	0	1	0	1	0
Apply U_T	0	0	0	1	0	1	0	0	0	1	1	1	0	0	0	0
Apply U_G	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1
Apply U_C	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0

Figura 4. Método proposto por Voss 1992 para a representação numérica de sequências de DNA
Fonte: Tiwari *et al.* 1997

Como demonstrado pela Figura 4, levando em consideração como o conjunto dos vetores sendo uma única matriz de quatro linhas e quantas colunas forem o número de bases da sequência, em cada coluna dessa matriz só será possível um único “1”, pelo motivo de não ser possível haver duas bases em uma mesma posição da sequência. Esse método binário é extramente útil, porém o fato de utilizar 4 vetores torna-o custoso computacionalmente quanto se trabalho com sequências contendo grandes quantidade de bases.

Há também outros métodos para a representação numérica de sequências de DNA, como por exemplo o método utilizado em Pessoa 2004, proposto por Cheever *et al.*1989, que representa a sequência no plano complexo, mapeando as base *A,T,C,G* respectivamente com +1,-1,+*i*,-*i*, onde *i* é a unidade imaginária (raiz quadrada de -1). A maior vantagem de se trabalhar com números complexos é que esses são geralmente a entrada e a saída da maioria das implementações de TFDs.

Basicamente, com o que vimos até aqui temos um processo semelhante ao que está ilustrado na Figura 5. Onde uma dada sequência de DNA é representada numericamente pelo método binário proposto por Voss 1992, e posteriormente é aplicado a TDF para cada um dos quatro vetores *Ux*, obtendo-se assim o espectro para cada vetor. Após realizar o somatório desses vetores, obtemos o espectro total referente a dada sequência de DNA. Quando plotamos o gráfico dos coeficientes obtidos com a TDF, atestamos a característica de periodicidade 3 de regiões codificadoras apontada por Fickett 1982, conforme demonstrado por Voss 1992 a possibilidade de visualizar essa característica aplicando a TDF em uma sequência de DNA.

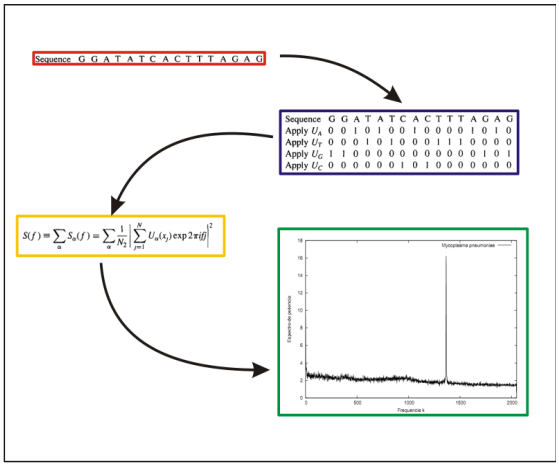

Figura 5. Geração do espectro com pico em *f*=1/3 a partir de uma sequência de DNA

Esse método descrito até agora não é capaz de encontrar regiões codificadoras em sequências de DNA, ele apenas demonstra a existência de uma ou mais regiões codificadoras em uma determinada sequencia de DNA. Porém, baseado no que foi mostrado, Tiwari *et al.* 1997 propõem um métado para a predição gênica. O método utiliza uma “janela” de tamanho *M*, nos experimentos chegou-se a conclusão que um tamanho interessante para a janela seria M = 391. Esse número foi o que produziu a menor quantidade de falsos negativos, e também, existem muitas poucas sequências e também *open reading frames* (ORFs) com tamanho inferior a 300 bases. O método vai “deslisando” a janela pela sequência, aplicando FTD sobre cada trecho e calculando os picos locais de cada um. Ao final, todos os espectros são unidos e considera-se região codificadoras as que apresentarem pico local maior ou igual a 4, conforme demonstrado pela Figura 6.

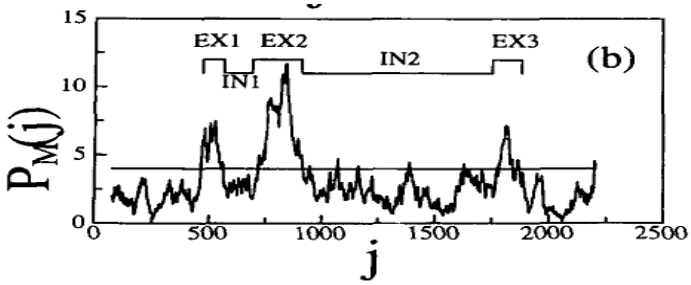

Figura 6. Espectro total da sequência de DNA do glóbulo β gerado pelo método proposto por Tiwari *et al.* 1997 **Fonte:** Tiwari *et al.* 1997

O método foi testado com uma grande quantidade de sequências de diferentes espécies (incluindo o genoma humano), somando um total de 5.5milhões de bases analisadas. Os resultados obtidos foram confrontados com o que se tem conhecido na literatura, chegando ao valor de 95% de acerto nas regiões codificadoras encontradas e 90% de acerto nas regiões não codificadoras. Alguns comparativos da predição realizada pelo método proposto por Tiware *et. al.* 1997 com as informações conhecidas pela literatura corrente, conforme mostrado pela Tabela 1.

Tabela 1. Comparativo entre a predição proposta em Tiwari *et al.* 1997 e informações existentes na literatura corrente. **Fonte:** Tiwari *et al.* 1997

Group	Species	G + C (%)	ORFs[a]	Genes[b]	ORFs with $P \geq 4$	Genes with $P \geq 4$
Fungus	*S.cerevisiae* III	38.6	216	54	198	51
Fungus	*S.cerevisiae* VIII	38.2	267	140	255	139
Insect virus	*A.californica*	40.7	154	51	137	49
Protozoa	*E.histolytica*[c]	34 2	26	26	26	26
Bacteria	*A.vinelandii*[d]	62.4	6	6	6	6
Bacteria	*H.influenzae*	38.2	1727	933	1667	927
Nematode	*C.elegans*	35 6	–	146	–	146
Mammal	Human[e]	51.2	–	24	–	24
Various	Globins[f]	49.6	–	15	–	15
Various	Actins[g]	36.8	15	15	15	15

Conforme citado no parágrafo anterior, o método proposto por Tiwari *et. al.* 1997 foi capaz de detectar regiões intrônicas e regiões exônicas com precisão de 90% e 95% respectivamente. Talvez não seja o valor ideal para se trabalhar com pesquisas, um método como precisão de 90% indica a produção de muitos falsos positivos. Porém, Twiari *et. al.* 1997 é um trabalho de grande relevância na biologia computacional, pois foi um dos pioneiros na utilização de processamento de sinais digitais para buscar informações relevantes em sequência de DNA. Além disso, o método apresentado por eles também foi capaz de gerar a predição (mesmo com algum grau de imprecisão) sem a necessidade de trabalhar com informações já conhecidas de códons ou de sequências de mesmas espécies, fazendo uso apenas da periodicidade 3 das regiões codificadoras, uma característica que é tida como universal. Sendo possível utilizá-lo para diferentes espécies.

RELATÓRIO DA IMPLEMENTAÇÃO

Conforme sugerido pela Profa. Hela Cristina, foi realizada uma implementação (código fonte em anexo ao resumo) relacionada ao tema e artigos estudos. No meu caso, a implementação reflete basicamente o que está representado pela Figura 5. Uma sequência de DNA é extraída de um arquivo no formato GenBank, a partir da sequência são gerados 4 vetores (conforme método proposto por Voss 1992). A Transformada de Fourier é aplicada para um dos quatro vetores e posteriormente realizado o somatório de cada um dos espectros para obter o espectro total da sequência conforme o método apresentado no decorrer do artigo, descrito em Tiwari *et al.* 1997. Por último, é gerado um arquivo um arquivo texto contendo com duas colunas, uma com a frequência e outro com o espectro de potência (parte real do coeficiente retornado pela TFD). O próprio programa chama o *gnuplot (www.gnuplot.info)* para plotar o gráfico e apresentá-lo ao usuário.

Alguns testes foram realizados e conseguimos obter os mesmos resultados descritos por Tiwari *et al.* 1997 e Pessoa 2004 mencionados ao longo desse resumo.

A implementação foi feita utilizando a linguagem C e compilada utilizando o *gcc* (http://gcc.gnu.org/) em um ambiente Linux utilizando a distribuição Fedora 15 (http://fedoraproject.org/) com arquitetura x86_64 e *kernel* do Linux na versão 2.6.38.8. As Transformadas de Fourier foram feitas utilizando a API FFTW na versão 3.2.2 (*Fastest Fourier Transform in the West*), também implementada em C e tida como uma das das implementações mais rápidas de FFT. Para compilar o programa é necessário "linkar" a biblioteca relativa a FFTW, como por exemplo: `gcc -o fft-jivago -lfftw3 -lm fft-jivago.c`

Após compilado, para executar o programa é necessário passar por parâmetro o arquivo no padrão GenBank contendo a sequência a ser analisada. Como por exemplo: `./fft-jivago sequences/AF135270.gb`. O resultado para essa chamada está representado pela Figura 7.

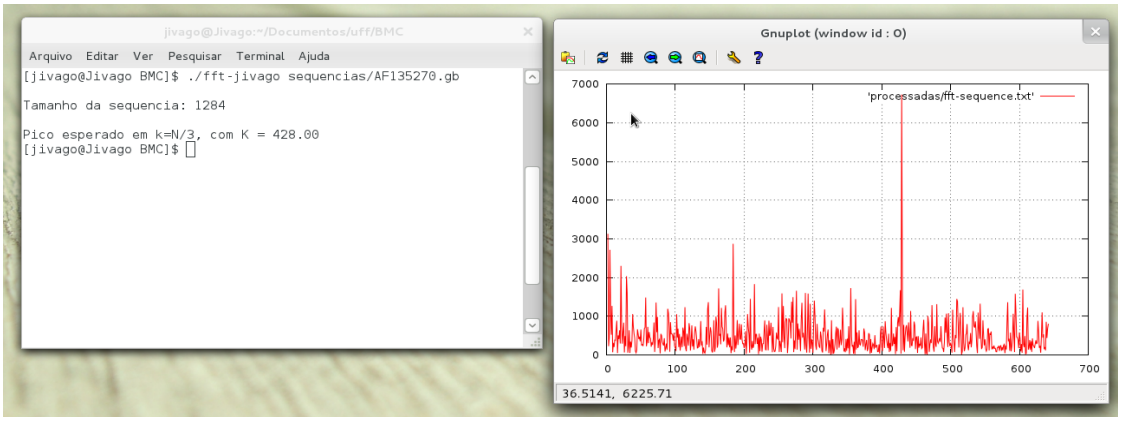

Figura 7. Pico no espectro da sequência gerado pela nossa implementação

Na Figura 7 vemos na direita o espectro total da sequência plotado no *gnuplot* após chamada do nosso programa. Nele podemos ver o pico esperado em f=1/3 para regiões codificadoras. Na esquerda temos a tela do terminal linux, mostrando a chamada do programa com o arquivo AF135270.gb sendo passado como parâmetro, e logo a baixo uma mensagem exibindo o tamanho total da sequência juntamente com o local esperado do pico f=1/3.

REFERÊNCIAS

Cheever, E., A., Searls, D., B., Karunaratne, W., Overton, G., C. **Using signal processing techniques for DNA sequence comparison. Bioengineering Conference.** 1989. 173-174.

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

Pessoa, S., L. **Análise da informação mútua em sequências de DNA homólogas**. Dissertação de Mestrado – Programa de Pós Graduação em Computação, Universidade Federal Fluminense, Niterói – RJ. 2004.

Tiwari S., Ramachandran S., Bhattacharya A., Bhattacharya S., Ramaswamy R. **Prediction of probable genes by Fourier analysis of genomic sequences**. Comput. Appl. Biosci. 1997; 13 (3) : 263-270

Voss., R., F.**Evolution of long-range fractal correlations and l/*f* noise in DNA base sequences**. Phys. Rev. Lett., 1992. 68, 3805-3808