ANÁLISE DE GENOMAS DE PROCARIOTES VIA FERRAMENTAS DE PROCESSAMENTO DIGITAL

Aluno: Clarice Dantas Silva
Orientador: Hélio Magalhães de Oliveira, Docteur

Departamento de Eletrônica e Sistemas, CTG, UFPE, 50.670-901,
Recife-PE, (81) 21268210, hmo@ufpe.br

A análise genômica vem recebendo grande atenção, particularmente após o seqüenciamento de muitos genomas, incluindo o genoma humano. O processamento de sinais genéticos envolve uma admirável quantidade de informação; isto representa um desafio na extração do conteúdo relevante destes sinais. A maioria dos métodos de análise de sinais genômicos para extração de atributos focaliza padrões de nucleotídeos nas seqüências de DNA. Novas ferramentas chamadas de códongramas e a^{2}gramas, foram recentemente introduzidas. Cada códon é representado por uma seqüência de três bases (c_1,c_2,c_3), em que c_i são nucleotídeos, $c_i \in N := \{A,T,C,G\}$, i=1,2,3. Estes são substituídos de acordo com: T→[11]; A→[-1-1]; G→[1-1]; C→[-11].

Operação 1 (*padding*). Primeiramente uma fita de DNA é convertida em um vetor DNA $\underline{g}$, acrescentando um número p de zeros à seqüência original, $p := -C \pmod 3$, em que C denota o comprimento do genoma. **Operação 2** (*reading frame*). Três seqüências de diferentes quadros de leitura (*rf*) podem ser geradas a partir de uma seqüência genômica g, denominadas $\dot{\underline{g}}, \ddot{\underline{g}}, \dddot{\underline{g}}$. Elas são compatíveis com o deslocamento cíclico da seqüência original. **Definição 1** (*Produto interno de seqüências de DNA*). Dadas duas seqüências de DNA de comprimento $3.\lceil C/3 \rceil$, $\underline{g}_1 = (\underline{c}^{1,1}\ \underline{c}^{1,2}\ \underline{c}^{1,3} \ldots \underline{c}^{1,\lceil C/3 \rceil})$ $\underline{g}_2 = (\underline{c}^{2,1}\ \underline{c}^{2,2}\ \underline{c}^{2,3} \ldots \underline{c}^{2,\lceil C/3 \rceil})$, o produto interno $\underline{g}_1 \bullet \underline{g}_2$ é definido por $\underline{g}_1 \bullet \underline{g}_2 := \sum_{j=1}^{\lceil C/3 \rceil} < \underline{c}^{1,j}, \underline{c}^{2,j} >$. □ Aqui, $\underline{c}^{i,j} = \left(c_1^{i,j}, c_2^{i,j}, c_3^{i,j}\right) \in N^3$ são 'códons' para i=1,2; j=1,2,…,$\lceil C/3 \rceil$. Se $\underline{g}_1 \bullet \underline{g}_2 = 0$ então os genes $\underline{g}_1$ e $\underline{g}_2$ são ditos DNA ortogonais. □ **Definição 2** (*produto vetorial entre seqüências de DNA*). Dadas as seqüências de DNA $\underline{g}_1$ e $\underline{g}_2$ de comprimento $3.\lceil C/3 \rceil$, o produto vetorial $\underline{g}_1 \otimes \underline{g}_2$ é dado por $\left(< \underline{c}^{1,1}, \underline{c}^{2,1} >, < \underline{c}^{1,2}, \underline{c}^{2,2} >, \ldots, < \underline{c}^{1,\lceil C/3 \rceil}, \underline{c}^{2,\lceil C/3 \rceil} > \right)$.

Vinte tipos de a^{2}gramas são definidos para um genoma, um para cada aminoácido. Os 'a^{2}gramas' proporcionam informação sobre os trechos da fita de DNA que potencialmente levem à síntese do amino ácido investigado. Um aplicativo com interface gráfica foi desenvolvido em MATLAB®. Estas ferramentas visuais para análise genômica podem ser usadas para pesquisa de padrões específicos de nucleotídeos. Entre tais padrões, o aplicativo inclui opções para determinação de: metgramas para estimar as posições de partida da codificação dos genes, localizadores de seqüência de Shine-Dalgarno (translação mRNA para proteínas), TATA Box (replicação DNA para mRNA). Estas ferramentas são atrativas na implementação de novas modalidades de classificação (*clustering*) de vírus e podem auxiliar na implementação de novos testes de diagnósticos para doenças genéticas, proporcionando um tipo de imagem médica de DNA.

Apoio: PIBIC/ UFPE/CNPq, projeto CNPq #306180.