II Encontro Acadêmico de Modelagem Computacional do LNCC
Petrópolis, Rio de Janeiro, Brasil
13 – 14 de Janeiro de 2009

INFERÊNCIA DIFUSA COMO SUPORTE À DESCOBERTA DE POSSÍVEIS SNPs EM SEQÜÊNCIAS DE cDNA

Wagner Arbex , Michel Eduardo Beleza Yamagishi , Marcos Vinícius G. Barbosa da Silva
Empresa Brasileira de Pesquisa Agropecuária, Rua Eugênio do Nascimento, 610, 36038-330, Juiz de Fora, MG
E-mail: arbex@cnpgl.embrapa.br, michel@cnptia.embrapa.br, marcos@@cnpgl.embrapa.br

Luiz Alfredo Vidal de Carvalho
Universidade Federal do Rio de Janeiro, Centro de Tecnologia - Bloco H-319, 21945-970, Rio de Janeiro, RJ
E-mail: alfredo@cos.ufrj.br

Diferenças pontuais entre pares de bases de diferentes seqüências alinhadas, são o tipo mais comum de variabilidade genética. Tais diferenças, conhecidas como polimorfismos base única (*single nucleotide polymorphisms* - SNPs), são importantes no estudo da variabilidade das espécies, pois podem provocar alterações funcionais ou fenotípicas, que, por sua vez, podem implicar em conseqüências evolutivas ou bioquímicas nos indivíduos das espécies. A descoberta de SNPs por algoritmos computacionais é uma prática bastante difundida e, nessa área, dois *scripts* se destacam pelo amplo uso: Polyphred [1] e Polybayes [2].

O Polyphred, analisa diretamente os sinais expressos no seqüenciamento do material genético e detecta SNPs a partir da variação dos sinais de fluorescência dos cromatogramas, procurando por reduções nas regiões do pico do sinal. Se for encontrada uma redução, onde uma segunda base foi detectada, então esse ponto é identificado como potencial heterozigoto. Após o alinhamento das seqüências (*reads*), as bases dessa seção transversal, que inclui *reads* e consenso, são comparadas. O Polybayes analisa as bases geradas a partir da "leitura" dos cromatogramas - feita por *base-calling* [3], que nomeia e atribui um valor de qualidade para cada base (*Phred quality score* - PQS) - e utiliza um algoritmo de inferência Bayesiana, que procura por seções transversais onde os *reads* alinhados apresentam bases diferentes entre si. O Polybayes considera o número de *reads* e, ainda, a taxa *a priori* de pontos polimórficos, como sendo $(\frac{1-0.003}{4})$, ou seja, um SNP para cada 333 pares de bases, dividido pelo número de possíveis diferentes bases - A, T, C ou G - em um ponto. Deve ser notado que, esses dois *scripts*, têm seus resultados influenciados pelo PQS, obtido durante a leitura dos cromatogramas.

Os referidos *scripts* trabalham com diferentes metodologias, sobre diferentes atributos, contudo, espera-se que apresentem resultados similares, ao tratarem um mesmo conjunto de seqüências, mas, não é incomum fornecerem resultados diferentes, o que produz incerteza na tomada de decisão, quando os resultados são discordantes.

O presente texto apresenta um modelo que se baseia em lógica difusa (*fuzzy logic*) para, a partir dos resultados do Polyphred e do Polybayes, auxiliar na tomada de decisão, no caso em que as informações sejam divergentes e, também, na confirmação de informações coincidentes. Ou seja, utiliza a lógica difusa para dar suporte à decisão, avaliando os resultados gerados por dois diferentes métodos e, ainda, incluindo, explicitamente, o PQS das bases do consenso, como um "valorizador" adicional, que reduz os efeitos específicos de cada um dos *scripts*.

A metodologia aqui apresentada não define nenhum limiar de "corte", no que se refere ao PQS, pois, o modelo de inferência difusa, automaticamente, elimina os pontos de baixa qualidade, não classificando-os como SNPs. Os critérios para a definição das variáveis linguísticas (conjuntos difusos), seus qualificadores e das funções de pertinência (expressões 2, 3 e 4), fundamentaram-se:

1. no índice atribuído pelo Polyphred (*Polyphred score* - PPS), que estabelece seis classes com intervalos *crisps*, variando de 1, que indica um $PPS \leq 49$ e um taxa de verdadeiros positivos de 1%, sendo improvável a existência de SNPs. Até 6, que indica $PPS \geq 99$ e uma taxa de verdadeiros positivos de 97%, sendo altamente provável a existência de SNPs, e, então, a variável linguística probabilidade foi definida nos termos: improvável (P_{IM}), pouco provável (P_{PP}), medianamente provável (P_{mP}), provável (P_{PR}), muito provável (P_{MP}) e altamente provável (P_{AP});

$$P_{IM}(x) = \begin{cases} 1 & x \leq 49 \\ \frac{59-x}{59-49} & 49 < x < 59 \\ 0 & x \geq 59 \end{cases} \quad P_{PP}(x) = \begin{cases} 0 & x \leq 25 \\ \frac{x-25}{50-25} & 25 < x < 50 \\ 1 & 50 \leq x \leq 69 \\ \frac{79-x}{79-69} & 69 < x < 79 \\ 0 & x \geq 79 \end{cases} \quad P_{mP}(x) = \begin{cases} 0 & x \leq 60 \\ \frac{x-60}{70-60} & 60 < x < 70 \\ 1 & 70 \leq x \leq 89 \\ \frac{91,5-x}{91,5-89} & 89 < x < 91,5 \\ 0 & x \geq 91,5 \end{cases} \tag{2}$$

$$P_{PR}(x) = \begin{cases} 0 & x \leq 80 \\ \frac{x-80}{90-80} & 80 < x < 90 \\ 1 & 90 \leq x \leq 94 \\ \frac{96-x}{96-94} & 94 < x < 96 \\ 0 & x \geq 96 \end{cases} \quad P_{MP}(x) = \begin{cases} 0 & x \leq 92,5 \\ \frac{x-92,5}{95-92,5} & 92,5 < x < 95 \\ 1 & 95 \leq x \leq 98 \\ \frac{99-x}{99-98} & 98 < x < 99 \\ 0 & x \geq 99 \end{cases} \quad P_{AP}(x) = \begin{cases} 0 & x \leq 96,5 \\ \frac{x-96,5}{99-96,5} & 96,5 < x < 99 \\ 1 & x \geq 99 \end{cases} \tag{3}$$

2. na qualidade das bases (PQS), que varia entre 4 e 60, separadas, pelo limiar $PQS = 20$, em duas classes de valores *crisps* e, então, a variável linguística qualidade foi definida nos termos: ruim (Q_R), boa (Q_B) e ótima (Q_O).

$$Q_R(x) = \begin{cases} 1 & x \leq 30 \\ \frac{30-x}{30-20} & 20 < x < 30 \\ 0 & x \geq 30 \end{cases} \quad Q_B(x) = \begin{cases} 0 & x \leq 20 \\ \frac{x-20}{30-20} & 20 < x < 30 \\ 1 & 30 \leq x \leq 40 \\ \frac{50-x}{50-40} & 40 < x < 50 \\ 0 & x \geq 50 \end{cases} \quad Q_O(x) = \begin{cases} 0 & x \leq 40 \\ \frac{x-40}{50-40} & 40 < x < 50 \\ 1 & x \geq 50 \end{cases} \tag{4}$$

Assim, no modelo de inferência aqui proposto, os valores discretos de entrada - os PPSs, encontrado pelo Polyphred e pelo seu equivalente no Polybayes, e o PQS - têm seus graus de pertinência estabelecidos pelas expressões 2, 3 e 4, que "disparam" regras difusas, cujo resultado é discretizado pelo método do "Centro do Máximo" (*Middle-of-Maxima* - MoM), visto que esse considera a ocorrência de múltiplos disparos de regras sobre uma mesma saída, "valorizando" essa saída. Desse modo, como resultado, determina-se um novo valor, mais apurado, indicativo da existência de polimorfismo, para cada SNP anteriormente identificado, onde foram considerados os valores iniciais dos PPSs e da PQS no ponto.

Palavras-chave: *Suporte à decisão com inferência difusa, Polimorfismo e Polimorfismo de base única.*

Referências

[1] Nickerson, D. A., Tobe, V. O. and Taylor, S. L.. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. Nucleic Acids Research, 25 (14): 2745-2751, 1997.

[2] Marth, G. T., Korf, I., Yandell, M. D., Teh, R. T., Gu, Z., Zakeri, H., Stitziel, N. O., Hillier, L., Kwok, P. Y. and Gish, W. R.. A general approach to single-nucleotide polymophism discovery. Nature Genetics, 23 (4): 452-456, 1999.

[3] Ewing, B., Hillier, L., Wendl, M. C. and Green, P.. Basecalling of automated sequencer traces using Phred (I). Genome Research, 8 175-185, 1998.

Sessão Técnica: Melhores Trabalhos Selecionados

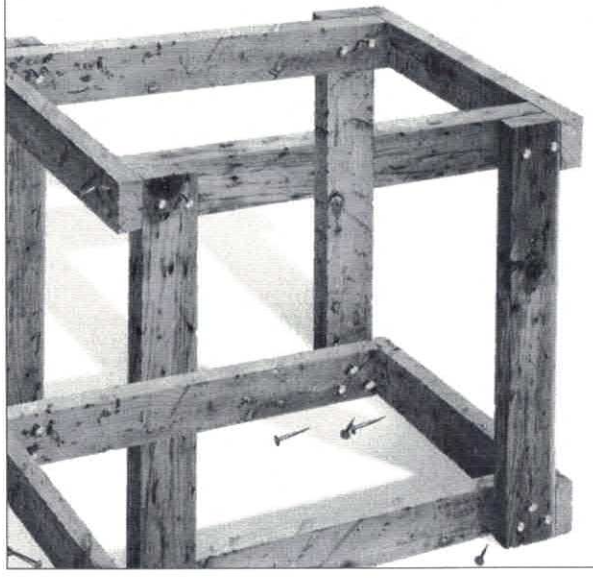
II ENCONTRO ACADÊMICO

MODELAGEM COMPUTACIONAL

Laboratório Nacional de Computação Científica

13 e 14 de Janeiro de 2009

Petrópolis - RJ



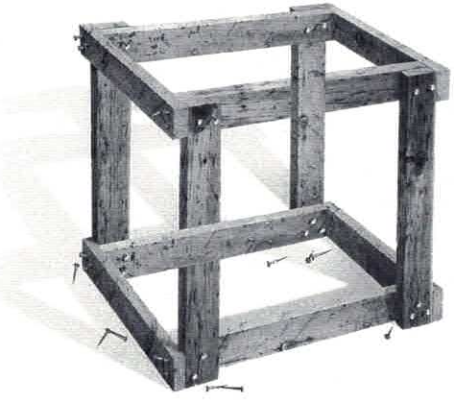
RESUMOS

II Encontro Acadêmico de Modelagem Computacional
do Laboratório Nacional de Computação Científica

IIEAMC-LNCC/MCT

Petrópolis, Rio de Janeiro, Brasil
13 – 14 de Janeiro de 2008

RESUMOS



Programação

13 de Janeiro de 2008 – Auditório A

HORÁRIO	EVENTO
09:00 - 10:00	**P1 – Palestra de Abertura** Eugenius Kashurewicz (FINEP)
10:00 - 10:15	**CO1 – Modelagem Multiescala - Técnica de Homogeneização** Riedson Baptista (LNCC)
10:15 - 10:30	**CO2 – Dimensão Dinâmica e Fractalidade** Marcelo Miranda Barros (LNCC)
10:30 - 10:50	**ST-P – Apresentação de Pôsteres**
11:00 - 11:30	**P2 – Modelagem Computacional e Verificações Experimentais em Dinâmicas das Estruturas** Flavio de Souza Barbosa (UFJF)
11:30 - 11:45	**CO3 – A Cellular Automata Framework for Real Time Fluid Animation** Sicilia Ferreira Ponse Pasini Judice (LNCC)
11:45 - 12:00	**CO4 – Evolução Diferencial para Otimização Restrita** Eduardo Krempser da Silva (LNCC)
12:00 - 12:30	**P3 – Jogos Educativos e Inovadores** Túlio Soria (UNESP)
12:30 - 13:30	Almoço
13:30 - 14:00	**P4 – Criação e Uso de Bases de Dados Secundárias** José Miguel Ortega (UFMG)
14:00 - 14:15	**CO5 – The Importance of Different Temperatures and Electrostatic Treatments on Molecular Dynamics Simulations** Priscila Vanessa Zabala Capriles Goliatt (LNCC)

P – Palestra
CO – Comunicação Oral
ST-P – Sessão Técnica-Pôster

13 de Janeiro de 2008 – Auditório A

HORÁRIO	EVENTO
14:15 - 14:30	**CO6 – Por que Predadores não Têm Efeitos Positivos sobre Presas?** José Carlos Lisboa Recarey Eiras (LNCC)
14:30 - 14:45	**CO7 – A Multidisciplinaridade da Computação Quântica - Desenvolvimento e Perspectivas** Amanda Castro Oliveira (LNCC)
14:45 - 15:00	**CO8 – Assimilação de Dados de Umidade do Solo Aplicado ao Modelo Hidrológico NOAH** Claudia Adam Ramos (LNCC)
15:00 - 15:30	**P5 – Modelagem Integrada da Atmosfera-Biosfera-Hidrosfera** Marcos Heil Costa (UFV)
15:30 - 15:50	**ST-P – Apresentação de Pôsteres**
16:00 - 16:15	**ST-CO1 – Processo de Busca do Parâmetro de Regularização Ótimo em um Problema de Restauração de Imagens** Dalmo Stutz (UERJ/IPRJ)
16:15 - 16:30	**ST-CO2 – A Performance Evaluation of a Parallel Implementation of the Inexact Jacobian-Free Newton-Krylov Method Using the PETSc Numerical Framework** Rafael Santos Coelho (UFES)
16:30 - 16:45	**ST-CO3 – A Nonlinear Subgrid Discontinuous Galerkin Method for Transport Problems** Natalia Arruda (LNCC)
16:45 - 17:00	**ST-CO4 – Modeling Heat Treatment of Steels by the Calorimeter Method** Cecilia Grisel Arone (KB Engineering S.R.L.)
17:00 - 17:15	**ST-CO5 – Inferência Difusa como Suporte à Descoberta de Possíveis SNPs em Seqüências de cDNA** Wagner Arbex (EMBRAPA)

P – Palestra
CO – Comunicação Oral
ST-P – Sessão Técnica-Pôster
ST-CO – Sessão Técnica-Comunicação Oral

Sumário