

MODELAGEM DIFUSA PARA SUPORTE À DECISÃO NA DESCOBERTA DE SNPs EM SEQUÊNCIAS DE cDNA

WAGNER ARBEX¹

LUIZ ALFREDO VIDAL DE CAVALHO²

MARCOS VINÍCIUS BARBOSA DA SILVA³

MICHEL EDUARDO BELEZA YAMAGISHI⁴

RESUMO: Diferenças pontuais entre pares de bases de diferentes sequências alinhadas, são o tipo mais comum de variabilidade genética. Tais diferenças, conhecidas como polimorfismos de base única (*single nucleotide polymorphisms* – SNPs), são importantes no estudo da variabilidade das espécies, pois podem provocar alterações funcionais ou fenotípicas, que podem implicar em consequências evolutivas ou bioquímicas nos indivíduos das espécies. A descoberta de SNPs por algoritmos computacionais é uma prática bastante difundida e o presente texto apresenta um modelo que se baseia em lógica difusa (*fuzzy logic*) para, a partir de resultados prévios, auxiliar na tomada de decisão, no caso em que as informações preliminares sejam divergentes, assim como, na confirmação de informações coincidentes.

PALAVRAS-CHAVE: modelagem difusa, inferência difusa, descoberta de conhecimento, polimorfismo de base única, variabilidade genética, fuzzyMorphic.pl.

FUZZY MODELLING TO AID DECISION IN THE IDENTIFICATION SNPs IN cDNA SEQUENCES

ABSTRACT: Differences between specific base pairs of different aligned sequences are the most common type of genetic variability. Such differences, known as single nucleotide polymorphisms (SNPs), are important in the study of the variability of the species, because they may cause functional or phenotypic changes, that can result in evolutionary or biochemical effects in the individuals of the species. The SNPs discovery by computer algorithms is a widespread practice and this paper show a fuzzy logic model based to aid in decision, from previous results, when the preliminary informations are divergent, as well as, in the confirmation of coincident informations.

KEY-WORDS: fuzzy modelling, fuzzy inference, knowledge discovery, single nucleotide polymorphism, genetic variability, fuzzyMorphic.pl.

1. INTRODUÇÃO

Uma das variações e particularidades do genoma da maioria das espécies são os chamados polimorfismos de base única (*single nucleotide polymorphisms* – SNPs), modificações de um único nucleotídeo, em uma dada sequência, quando comparada a outra. Ou seja, SNPs são pares de bases em uma única posição do DNA genômico, que se apresentam com diferentes alternativas nas sequências (Figura 1) – isto é, alelos – e podem ser encontrados no genoma de indivíduos normais em algumas populações ou grupos de indivíduos.

¹ Doutor em Engenharia de Sistemas e Computação, Empresa Brasileira de Pesquisa Agropecuária, E-mail: arbex@cnpgl.embrapa.br

² Doutor em Engenharia de Sistemas e Computação, Universidade Federal do Rio de Janeiro, E-mail: alfredo@cos.ufrj.br

³ Doutor em Genética e Melhoramento, Empresa Brasileira de Pesquisa Agropecuária, E-mail: marcos@cnpgl.embrapa.br

⁴ Doutor em Matemática Aplicada, Empresa Brasileira de Pesquisa Agropecuária, E-mail: michel@cnptia.embrapa.br

... GGG <u>AAA</u> CTCCAG...	... GGG <u>AAA</u> CTCCAG...	... GGG <u>AAA</u> CTCCAG...
... GGG AAA CTCCAG...	... GGG ACA CTCCAG...	... GGG ATA CTCCAG...
... GGG AAA CTCCAG...	... GGG AAA CTCCAG...	... GGG AAA CTCCAG...
... GGG AGA CTCCAG...	... GGG AGA CTCCAG...	... GGG ACA CTCCAG...
... GGG AGA CTCCAG...	... GGG AGA CTCCAG...	... GGG AGA CTCCAG...

Figura 1: Exemplos hipotéticos de SNPs bi, tri e tetra-alelicos, respectivamente. A primeira linha, em negrito, representa a sequência consenso e as bases sublinhadas, os polimorfismos.

A maior parte do genoma entre os indivíduos de uma mesma espécie é idêntica, porém, existe a variabilidade genética, que consiste na alteração nas sequências de bases ao longo do DNA, ocorrem por substituição, ausência ou duplicação de bases e, os SNPs, são o tipo mais comum de variabilidade genética (HapMap, 2003).

O que difere um indivíduo dos demais da sua espécie é o código genético, isto é, as sequências de nucleotídeos que formam as moléculas e sequências de DNA, RNA e proteínas, que, por sua vez, interagem e formam as células, as quais, por sua vez, formam os tecidos, os órgãos, até que, finalmente, formam os indivíduos. Ou seja, as diferenças se iniciam na ordem em que os nucleotídeos se apresentam para, após um complexo processo que envolve transcrição e tradução, originarem as proteínas. Essa é a importância dos SNPs, pois, a alteração de um único nucleotídeo em uma dada sequência, pode alterar a produção de uma certa proteína, e são importantes no estudo da variabilidade das espécies, uma vez que podem provocar alterações funcionais ou fenotípicas, que podem implicar em consequências evolutivas ou bioquímicas nos indivíduos em que os SNPs se manifestam.

2. OBJETIVO

Apresentar e discutir um modelo de inferência difusa para suporte à decisão na descoberta de SNPs em sequências de cDNA, permitindo sua implementação por meio de um modelo computacional, fundamentado em aprendizado de máquina para descoberta de conhecimento em bases de dados (*knowledge discovery in database – KDD*).

3. MATERIAIS E MÉTODOS

A discussão do modelo proposto passa pela obtenção e montagem de sequências de cDNA, a busca de SNPs nessas sequências e, ainda, a modelagem computacional para suporte a decisão por meio de inferência difusa, que são abordadas na presente seção.

A partir da montagem de sequências-consenso, pelo alinhamento de sequências obtidas por algum procedimento de análise e interpretação do DNA genômico, se inicia a investigação em sequências de cDNA e, para esse trabalho, o procedimento utilizado para tal foi a geração de sequências expressas identificadas (*expressed sequence tags – EST*).

Também conhecidas como sequências derivadas de transcritos, as ESTs, são medições dos níveis de mRNA (Lesk, 2005), que é a porção extraída do RNA transcrito e traduzida em proteína, e geram sequências curtas de cDNA que descrevem padrões de transcrição de genes. Os procedimentos de sequenciamento por meio de ESTs, baseiam-se em “capturar” essas sequências de nucleotídeos de regiões expressas do genoma que correspondem a pequenas sequências com, aproximadamente, 200 e 500 pares de bases (Adams et al., 1991). Assim, torna-se necessário agrupar (*cluster*) e alinhar as sequências obtidas que correspondem a fragmentos de um mesmo gene, para que seja possível montar sequências mais extensas.

A descoberta de SNPs por algoritmos computacionais é uma prática bastante difundida e os programas Polyphred e Polybayes se destacam pelo amplo uso. O Polyphred, analisa os sinais expressos no sequenciamento 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 onde uma segunda base foi detectada (Nickerson et al., 1997). Por sua vez, o Polybayes,

analisa as bases geradas a partir da “leitura” dos cromatogramas (Ewing et al., 1998) – 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 as sequências alinhadas apresentam bases diferentes entre si (Marth et al., 1999).

Os referidos programas implementam diferentes avaliações, sobre diferentes atributos, contudo, espera-se que apresentem resultados similares, ao tratarem um mesmo conjunto de sequências, mas, não é incomum fornecerem resultados diferentes, o que produz incerteza na tomada de decisão. Além disso, deve ser notado que, esses dois programas, têm seus resultados influenciados pelo PQS, obtido durante a leitura dos cromatogramas.

Modelos de inferência difusa são adequados para representar a informação imprecisa, que pode ser expressa por um conjunto de regras linguísticas e, caso exista a possibilidade de que os operadores sejam organizados como um conjunto de regras da forma

se ANTECEDENTE então CONSEQUENTE

logo, o raciocínio subjetivo pode ser construído em um algoritmo computacionalmente executável (Tanscheit, 2007), com capacidade de classificar, de modo impreciso, as variáveis que participam dos termos antecedentes e consequentes das regras, em conceitos qualitativos, e não quantitativos, o que representa a idéia de variável linguística (Almeida et al., 2005).

Trabalhar com valores incertos possibilita a modelagem de sistemas complexos, mesmo que reduza a precisão do resultado, o que não retira a credibilidade. Se as incertezas, quando consideradas isoladamente, são indesejáveis, quando associadas a outras características dos sistemas a serem modelados, em geral, permitem a redução da complexidade do sistema e aumentam a credibilidade dos resultados obtidos (Klir et al., 1995).

A subjetividade no raciocínio, sendo transmitida e compreendida entre interlocutores, é expressa em “termos e variáveis linguísticas” (Zadeh, 1973), mas não é expressa pela lógica clássica ou qualquer abordagem matemática tradicional. Por exemplo, o uso de adjetivos que representam imprecisão ou incerteza, tais como, alto, baixo ou, ainda, relações como, o conjunto das pessoas altas, não podem ser expressos por essas abordagens, a menos que seja definido, por exemplo, a altura, a partir da qual, uma pessoa pode ser considerada alta.

4. MODELO PROPOSTO

Cada ponto possivelmente polimórfico identificado pelo Polyphred ou pelo Polybayes, tem, entre outros atributos, a sua probabilidade estimada por cada um dos programas e o seu valor da qualidade da base – *Phrap quality* (PQ) – na sequência-consenso, que não é considerada diretamente por esses aplicativos.

Com os resultados do Polyphred e do Polybayes, o modelo de inferência auxilia a tomada de decisão, no caso em que as informações sejam divergentes e, também, na confirmação de informações coincidentes, avaliando esses resultados e inclui o PQ como um “valorizador” adicional, que reduz os efeitos específicos de cada um dos programas e amplia as possibilidades de investigação. No caso, PQ é utilizado na análise como apoio à decisão, então, aos dados prévios sobre a possibilidade de o ponto vir a ser um SNP, acrescenta-se a sua qualidade, com o objetivo de se estabelecer uma das três possibilidades: a confirmação de o ponto ser um SNP (SNP confirmado – SNP_C), a eliminação dessa possibilidade (SNP descartado – SNP_D) ou uma situação sem a confirmação dessa possibilidade, mas também sem elementos conclusivos para seu descarte (SNP não confirmado – SNP_{NC}).

No modelo proposto, as funções de pertinência adotadas foram baseadas:

1. No *Polyphred score* (PPS⁵), que estabelece seis classes com intervalos precisos

⁵ Ainda que o PPS estabeleça seis classes, essas classes são definidas a partir da probabilidade, entre 0% e 100%, do ponto vir a ser um SNP, o que coincide com a medida de probabilidade do Polybayes. Assim, a denominação PPS, apesar de referir-se ao Polyphred, está sendo utilizada para ambos, uma vez que o

(Nickerson et al., 1997), variando de 1, que indica um PPS ≥ 99 e uma taxa de verdadeiros positivos de 97%, sendo provável a existência de SNPs; até 6, que indica PPS ≤ 49 e uma taxa de verdadeiros positivos de 1%, sendo improvável a existência de SNPs. Assim, a sua função de pertinência foi definida pela variável linguística *probabilidade*, com os termos: improvável (P_{IM}), pouco provável (P_{PP}), medianamente provável (P_{mP}), provável (P_{PR}), muito provável (P_{MP}) e altamente provável (P_{AP});

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

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

2. No PQ que varia entre 4 e 90 e é separado, pelo limiar⁶ PQ = 20, em duas classes de valores, sendo sua função de pertinência definida como a variável linguística *qualidade*, nos termos: ruim (Q_R), boa (Q_B) e ótima (Q_O).

$$Q_R(x) = \begin{cases} 1 & x \leq 20 \\ \frac{30-x}{30-20} & 20 < x < 30 \\ 0 & x \geq 30 \end{cases} \quad Q_B(x) = \begin{cases} 0 & x \leq 15 \\ \frac{x-15}{30-15} & 15 < x < 30 \\ 1 & 30 \leq x \leq 40 \\ \frac{70-x}{70-40} & 40 < x < 70 \\ 0 & x \geq 70 \end{cases} \quad Q_O(x) = \begin{cases} 0 & x \leq 40 \\ \frac{x-40}{50-40} & 40 < x < 50 \\ 0 & x \geq 50 \end{cases}$$

Essas variáveis são utilizadas em trinta e seis regras de inferência, divididas em dois conjuntos, que possuem a respectiva metade para avaliar *qualidade* e *probabilidade*, segundo Polyphred e Polybayes, e que relacionam as combinações possíveis dos termos linguísticos:

Regra 1: se Q_R e P_{IM} então SNP_D

Regra 2: se Q_R e P_{PP} então SNP_D

:

Regra 18: se Q_O e P_{AP} então SNP_C

Cada um desses dois conjuntos inferem as decisões representadas na Tabela 1.

	P_{IM}	P_{PP}	P_{mP}	P_{PR}	P_{MP}	P_{AP}
Q_R	SNP_D	SNP_D	SNP_D	SNP_D	SNP_D	SNP_D
Q_B	SNP_D	SNP_D	SNP_{NC}	SNP_{NC}	SNP_C	SNP_C
Q_O	SNP_D	SNP_D	SNP_{NC}	SNP_{NC}	SNP_C	SNP_C

Tabela 1: Decisões inferidas pelo modelo a partir da avaliação de *qualidade* e *probabilidade*.

5. RESULTADOS E DISCUSSÃO

Os termos e variáveis linguísticas aumentam a complexidade de um sistema de computação frente à capacidade desses sistemas trabalharem com números ou outros valores exatos, discretos, muitas vezes, excludentes e o modelo proposto, definido a partir de técnicas de aprendizado de máquina, substitui, por meio de inferência difusa, medidas de probabilidade,

Polybayes possui uma exata medida equivalente.

⁶ O limiar de qualidade estabelece a divisão, a priori, entre as bases consideradas ruins, PQS < 20, das demais e foi definido a partir do critério Phred 20, visto que o PQ é calculado com base no PQS.

determinadas pelo Polyphred e Polybayes, contínuas no intervalo $[0, 1]$ e associadas à possibilidade de um ponto vir a ser um SNP, por um outro atributo, que permite decidir sobre a classe de cada ponto – SNP_C , SNP_D ou SNP_{NC} – e essa decisão somente pode ser tomada devido à possibilidade do modelo de inferência difusa refletir o raciocínio subjetivo a partir de regras de inferência. Esse modelo, além da fundamentação apresentada, foi determinado pelas características dos resultados obtidos com ferramentas usuais de identificação de SNPs em que critérios fixos e precisos – como o PPS - não são adequados, quando a são obtidos resultados próximos à divisão de classes ou quando essas ferramentas apresentam resultados conflitantes.

Uma maior discussão sobre o modelo proposto, pode ser encontrada em Arbex (2009), juntamente com a descrição da ferramenta fuzzyMorphic.pl, utilizada para o desenvolvimento e a implementação do mesmo.

6. CONCLUSÃO

Fundamentado na lógica difusa, foi elaborado um modelo de inferência difusa para suporte à decisão, pela a análise de resultados prévios e frente a características de incerteza, que são comuns em processos decisórios.

7. REFERÊNCIAS

- ADAMS, M. D.; KELLEY, J. M.; GOCAYNE, J. D.; et al. Complementary DNA sequencing: expressed sequence tags and human genome project, **Science**, v. 252, n. 5013, p. 1651-1656, jun. 1991.
- ALMEIDA, P. E. M.; EVSUKOFF, A. G. Sistemas fuzzy. **In: REZENDE, S. O. (ed.). Sistemas inteligentes: fundamentos e aplicações.** Barueri: Manole, 2005. p. 169–202.
- ARBEX, W. **Modelos computacionais para identificação de informação genômica associada à resistência ao carrapato bovino**, 2009. Tese de doutorado. Universidade Federal do Rio de Janeiro. 200 p.
- EWING, B.; HILLIER, L.; WENDL, M. C.; et al. Base-calling of automated sequencer traces using Phred (I): Accuracy assessment, *Genome Research*, v. 8, p. 175-185, 1998.
- HAPMAP The International HapMap Project, **Nature**, v. 426, n. 6968, p. 789-796, dec. 2003. (The International HapMap Consortium)
- KLIR, G. J.; YUAN, B. Fuzzy sets and fuzzy logic: theory and applications. Upper Saddle River: Prentice Hall, 1995. 592 p.
- LESK, A. M. **Introduction to bioinformatics.** New York: Oxford University Press, 2nd ed, 2005. 378 p.
- MARTH, G. T.; KORF, I.; YANDELL, M. D.; et al. A general approach to single-nucleotide polymorphism discovery, **Nature Genetics**, v. 23, p. 452-456, dec. 1999.
- NICKERSON, D. A.; TOBE, V. O.; TAYLOR, S. L. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing, **Nucl. Acids Res.**, v. 25, n. 14, p. 2745-2751, 1997.
- TANSCHKEIT, R. Sistemas fuzzy. **In: OLIVEIRA JR., H. A. (ed.). Inteligência computacional: aplicada à administração, economia e engenharia em Matlab.** São Paulo: Thomson Learning, 2007, pp. 229-264.
- ZADEH, L. A. Outline of a new approach to the analysis of complex systems and decision processes, **IEEE Trans. on Systems, Man, and Cybernetics**, v. SMC-3, p. 28-44, 1973.