X Congresso Brasileiro de Inteligência Computacional (CBIC'2011), 8 a 11 de Novembro de 2011, Fortaleza, Ceará

SISTEMA *FUZZY* APLICADO À BIOINFORMÁTICA NA TOMADA DE DECISÃO PARA IDENTIFICAÇÃO DE SNPs

Wagner Arbex, Marcos Vinícius Barbosa da Silva
Empresa Brasileira de Pesquisa Agropecuária
{arbex,marcos}@cnpgl.embrapa.br

Fabrízzio Condé de Oliveira
Universidade Salgado de Oliveira
fabrizzioconde@gmail.com

Luís Alfredo Vidal de Carvalho
Universidade Federal do Rio de Janeiro
LuisAlfredo@ufrj.br

Resumo – A investigação de polimorfismos de base única necessita de ferramentas de bioinformática que devem ser aplicadas a diferentes casos, com capacidade para analisar seqüências de diferentes fontes, níveis de cobertura e que consigam medidas confiáveis. Essas ferramentas trabalham com diferentes metodologias, sobre diferentes atributos, contudo, espera-se que apresentem resultados similares, ao tratarem um mesmo conjunto de dados, mas, não é incomum fornecerem resultados diferentes, o que produz incerteza na tomada de decisão, quando os resultados são discordantes. Esse texto mostra um sistema *fuzzy* que implementa um modelo de inferência para apoio à decisão aplicado à bioinformática, especificamente, na identificação de polimorfismos de base única, a partir de resultados oriundos de duas outras ferramentas de descoberta de tais polimorfismos.

Palavras-chave – Sistema de inferência fuzzy, tomada de decisão, sistema fuzzy, lógica fuzzy, polimorfismo de base única, bioinformática

Abstract – Research involving the discovery of single nucleotide polymorphisms (SNPs) requires bioinformatics tools to be applied to different cases, with the ability to analyze "reads" from different sources, levels of coverage and to establish reliable measures. These tools work with different methodologies on different attributes, however, it is expected similar results, even when dealing with a same data set, but it's not unusual to provide different results, which leads to uncertainty in decision making, when the results are discordant. This papes shows a fuzzy inference system that implements a fuzzy inference model decision support applied to bioinformatics, specifically, in the identification of SNPs, based on results from two other SNPs discovery tools

Keywords – Fuzzy inference system, decision maker, fuzzy system, fuzzy logics, single nucleotide polymorphisms, bioinformatics.

1 INTRODUÇÃO

As tecnologias de geração de dados para a biologia molecular desafiam o desenvolvimento de sistemas de computação adequados e requerem ferramentas de bioinformática precisas para a análise de tais dados. Sob tais aspectos, o aprendizado de máquina mostra-se como uma alternativa promissora [1], para a descoberta de conhecimento em bases de dados genômicos, com o uso de técnicas de mineração de dados entre outros recursos da IA.

O objetivo desse texto é o de apresentar um modelo matemático e computacional para tomada de decisão, desenvolvido e implementado com o fuzzyMorphic.pl [2], aplicado à investigação de polimorfismos de base única (*single nucleotide polymorphisms* - SNPs) em seqüências expressas de cDNA, que utiliza-se de lógica *fuzzy* para a implementação de um sistema de inferência, auxiliar à tomada de decisão, partindo de resultados prévios, obtidos por diferentes ferramentas de descoberta de SNPs e que apresentam resultados possivelmente conflitantes. O modelo é aplicado para auxiliar na tomada de decisão, nos casos em que as informações sejam divergentes e, também, na confirmação de informações coincidentes.

O fuzzyMorphic.pl é uma plataforma para desenvolvimento de *fuzzy inference systems* (FISs) que permite, por exemplo, a implementação de modelos computacionais capazes de processar informações imprecisas e qualitativas sendo, portanto, adequados em situações de tomada de decisão [3]. Suas características de implementação, bem como, seus recursos de descrição dos modelos, permitem o desenvolvimento de FISs para variados problemas e modelos baseados em inferência *fuzzy*, para os quais seja possível, para a etapa de fuzzificação, representar as funções de pertinência sobre formatos de conjuntos trapezoidais; para a implementação da máquina de inferência, utilizar os modelos de Mamdani ou de Larsen; para a defuzzificação, representar a função de saída sobre formatos de conjuntos trapezoidais e utilizar o centro dos máximos como método de defuzzificação.

2 POLIMORFISMOS DE BASE ÚNICA

Os projetos de seqüenciamento de genomas revelaram que esses possuem mais variações e maior complexidade do que inicialmente previa-se. Uma das variações e particularidades dos genomas são SNPs, isto é, pares de bases em uma única posição no DNA genômico, que se apresentam com diferentes alternativas nas seqüências [4] e podem ser encontrados no genoma de indivíduos isoladamente ou em grupos de indivíduos, em alguma população (Figura 1).

Figura 1: Exemplos hipotéticos de SNPs bi, tri e tetra-alélicos, respectivamente. A primeira linha, em negrito, representa a seqüência consenso e as bases sublinhadas, os SNPs. Na prática, a ocorrência de SNPs bi-alélicos não é somente mais comum, mas, quase absoluta em relação as demais formas [5].

A individualidade é conseqüência da expressão do código genético, ou seja, em sua essência, as seqüências de nucleotídeos formam as moléculas e seqüências de DNA, RNA e proteínas, que, por sua vez, interagem e formam as células, as quais também, interagem e formam os tecidos, os órgãos, até que, finalmente, formam os indivíduos. Essa é a importância dos SNPs, pois, em síntese, a alteração de um único nucleotídeo, uma única base, em uma dada seqüência, pode alterar a formação de proteínas e o conjunto dessas alterações pode provocar variações nas características dos indivíduos.

3 INFERÊNCIA DIFUSA COMO SUPORTE À DECISÃO

A subjetividade no raciocínio em geral, utilizada no cotidiano, sendo transmitida e perfeitamente compreendida entre interlocutores, é expressa em "termos e variáveis lingüísticas" [6] e não é expressa sob a lógica clássica ou qualquer abordagem matemática tradicional. O uso de, por exemplo, adjetivos comuns que representam imprecisão ou incerteza, tais como, *alto*, *baixo* ou, relações e agrupamentos, como, *conjunto das pessoas altas*, não podem ser expressos por essas abordagens, a menos que seja definido, com exatidão, o conceito ou o valor que determine a altura, a partir da qual, uma pessoa pode ser considera alta.

Os termos e variáveis lingüísticas aumentam a complexidade dos sistemas computacionais frente à capacidade de trabalharem com números, valores exatos, discretos e, por vezes, excludentes, o que sugere a idéia de que, trabalhar com valores incertos, possibilita a modelagem de sistemas complexos, mesmo que se reduza a precisão do resultado, mas não retira a credibilidade. Se as incertezas, quando consideradas isoladamente, são indesejáveis, quando associadas a outras características, em geral, permitem a redução da complexidade do sistema e aumentam a credibilidade dos resultados obtidos [7].

As abordagens clássicas são falhas para valores limítrofes e, portanto, resultados matemática e logicamente precisos, porém, questionáveis, podem ser encontrados. Por exemplo, o *Polyphred score* (*PPS*) estabelece seis classes com intervalos precisos (Tabela 1) [8] e, supondo que fossem determinados os *scores* 70 e 89 para dois pontos, então, para ambos, seria considerada a taxa de 35% de verdadeiros positivos na decisão desses pontos virem a ser SNPs (Classe 4).

Tabela 1: Classes definidas pelo PPS [8].

Classe	PPS	Taxa de verdadeiros positivos
1	99	97%
2	95 - 98	75%
3	90 - 94	62%
4	70 - 89	35%
5	50 - 69	11%
6	0 - 49	1%

Essa decisão, lógica e matematicamente precisa, pode ser questionada devido à subjetividade que a envolve, visto que, 70 e 89, se encontram nos extremos da classe a qual pertencem e, portanto, muito próximos de diferentes interpretações. Todavia, as abordagens clássicas da lógica e da matemática não possuem as ferramentas necessárias para tratar valores limítrofes, imprecisão ou incerteza. Um valor limítrofe acarretará dúvidas na "decisão" de o ponto ser, ou não, considerado polimórfico, o que sugere um FIS para o tratamento dessa incerteza.

O problema de valores limítrofes, em geral, não é tão simples quanto parece, do contrário, as abordagens clássicas poderiam facilmente resolvê-lo, mas, ao aproximar-se do raciocínio subjetivo para a interpretação e a extração de uma resposta, uma decisão, torna-se complexo e a aparente simplicidade é conferida pela modelagem por lógica *fuzzy* e seu embasamento na teoria dos conjuntos *fuzzy*. A subjetividade intrínseca ao raciocínio trata situações complexas, mediante imprecisão, incerteza ou aproximação e, então, são utilizados "operadores humanos", também de natureza imprecisa, que são expressos por termos ou variáveis lingüísticas, o que, em geral, não permite uma solução em termos exatos, mas, pode propor uma classificação, agrupamento ou agregação qualitativa em categorias ou possíveis conjuntos de soluções [3].

A lógica *fuzzy* e a sua teoria de conjuntos são adequados para representar, em termos matemáticos, a informação imprecisa, que pode ser expressa por um conjunto de regras lingüísticas e, caso exista a possibilidade de que os operadores humanos sejam organizados como um conjunto de regras da forma *se ANTECEDENTE então CONSEQÜENTE* logo, o raciocínio subjetivo pode ser construído em um algoritmo computacionalmente executável [9] com capacidade de classificar, de modo impreciso, as variáveis que participam dos termos antecedentes e conseqüentes das regras, em conceitos qualitativos, e não quantitativos, o que representa a idéia de variável lingüística [3]. Assim, como sistemas capazes de processar de forma eficiente informações imprecisas e qualitativas, os modelos de inferência difusa são adequados em situações que exigem tomadas de decisão [3].

4 DESCRIÇÃO DO MODELO E DO SID PARA IDENTIFICAÇÃO DE SNPs

Em geral, as etapas de um FIS são: a fuzzificação, que converte os dados "precisos" (*crisps*) de entrada em valores difusos; a inferência, propriamente dita; e a defuzzificação, que converte os resultados difusos em grandezas numéricas precisas. No modelo proposto, consideram-se como valores de entrada, as probabilidades, previamente determinadas, de o ponto vir a ser um SNP e o valor de qualidade do ponto na seqüência consenso. Os *Casos 1* e *2* serão utilizados ao longo do texto para demonstrar o modelo, assumindo, para o *Caso 1*, 99% e 96%, quanto as probabilidades e 43 de qualidade e, para o *Caso 2*, os valores são, respectivamente, 94%, zero e 50.

Esses casos são parte de um estudo realizado no projeto "Modelos computacionais para identificação de informação genômica associada à resistência ao carrapato bovino" [10], que desenvolveu, implementou e executou o FIS em discussão, sobre 4072 seqüências expressas relacionadas à expressão de resistência de bovinos à ação do "carrapato do boi", na busca de informações genômicas, em específico, SNPs que estivessem associados à resistência dos bovinos a esses ácaros.

Os valores das probabilidades, que, a princípio, deveriam ser classificados diretamente na Tabela 1, em uma tentativa de se identificar um SNP, foram obtidos com o uso dos programas Polyphred [11] e Polybayes [12]. 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* [13], 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.

Esses programas possuem diferentes métodos para a obtenção de seus resultados e podem apresentar valores muito conflitantes, como o exemplificado no *Caso 2*. Assim, uma comparação simples desses resultados com a Tabela 1, pode levantar ainda mais dúvidas, quando o que se buscava era uma resposta "exata" e, além disso, deve ser notado que, esses dois programas, têm seus resultados influenciados pelo *PQS*, obtido durante a leitura dos cromatogramas.

4.1 FUZZIFICAÇÃO

Avalia-se um valor de entrada por sua "função de pertinência", o que determina um "grau de pertinência" (GP) do valor para a sua função e as funções de pertinência adotadas foram baseadas:

1. no *PPS* (Tabela 1), com a função de pertinência definida pela variável lingüística *probabilidade*, com os termos (Expressões 1 e 2): *improvável* (P_{IM}), *pouco provável* (P_{PP}), *medianamente provável* (P_{mP}), *provável* (P_{PR}), *muito provável* (P_{MP}) e *altamente provável* (P_{AP});

2. na qualidade das bases do consenso – o *PQS* – que varia entre 4 e 90 e sua função de pertinência define a variável lingüística *qualidade*, nos termos (Expressões 3): *ruim* (Q_R), *boa* (Q_B) e *ótima* (Q_O).

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

$$P_{PR}(x) = \begin{cases} 0 & x \leq 80 \\ \frac{x-80}{90-80} & 80 < x < 90 \\ 1 & 90 \leq x \leq 94 \\ \frac{96-x}{96-94} & 94 < x < 96 \\ 0 & x \geq 96 \end{cases} \quad P_{MP}(x) = \begin{cases} 0 & x \leq 92,5 \\ \frac{x-92,5}{95-92,5} & 92,5 < x < 95 \\ 1 & 95 \leq x \leq 98 \\ \frac{99-x}{99-98} & 98 < x < 99 \\ 0 & x \geq 99 \end{cases} \quad P_{AP}(x) = \begin{cases} 0 & x \leq 96,5 \\ \frac{x-96,5}{99-96,5} & 96,5 < x < 99 \\ 1 & x \geq 99 \end{cases} \tag{2}$$

$$Q_R(x) = \begin{cases} 1 & x \leq 20 \\ \frac{30-x}{30-20} & 20 < x < 30 \\ 0 & x \geq 30 \end{cases} \quad Q_B(x) = \begin{cases} 0 & x \leq 15 \\ \frac{x-15}{30-15} & 15 < x < 30 \\ 1 & 30 \leq x \leq 40 \\ \frac{70-x}{70-40} & 40 < x < 70 \\ 0 & x \geq 70 \end{cases} \quad Q_O(x) = \begin{cases} 0 & x \leq 40 \\ \frac{x-40}{50-40} & 40 < x < 50 \\ 1 & x \geq 50 \end{cases} \tag{3}$$

Os resultados da fuzzificação para o *Caso 1*, $PPS_1 = 99$, $PPS_2 = 96$ e $PQS = 43$, em suas respectivas funções de pertinência, podem ser vistos nas Tabelas 2 e 3 e as Figuras 2 e 3 representam graficamente seus conjuntos difusos e, para o *Caso 2*, o resultado da fuzzificação para $PPS_1 = 94$, $PPS_2 = 0$ e $PQS = 50$, pode ser visto nas Tabelas 4 e 5 com as representações nas Figuras 4 e 5.

Tabela 2: GPs para a variável *probabilidade*, para o *Caso 1*.

	PPS_1	PPS_2
Improvável	0	0
Pouco provável	0	0
Medianamente provável	0	0
Provável	0	0
Muito Provável	0	1
Altamente provável	1	0

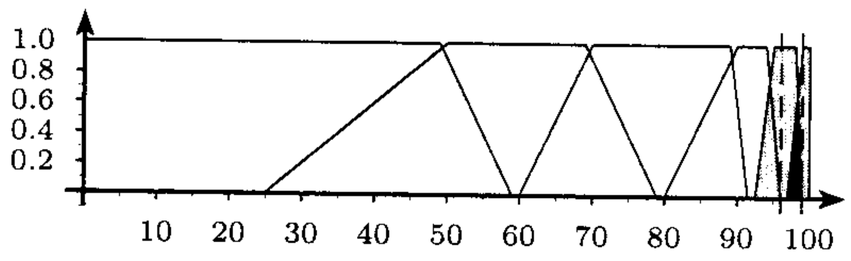

Figura 2: Fuzzificação para a variável *probabilidade*, no *Caso 1*.

Tabela 3: GPs para a variável *qualidade*, para a *Caso 1*.

	PQS
Ruim	0
Boa	0,9
Ótimo	0,3

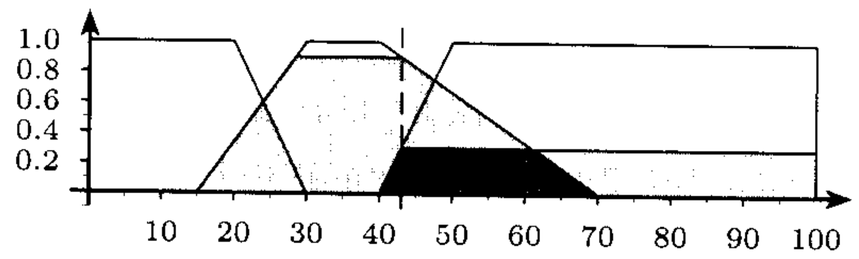

Figura 3: Fuzzificação para a variável *qualidade*, no *Caso 1*.

Tabela 4: GPs para a variável *probabilidade*, para o *Caso 2*.

	PPS_1	PPS_2
Improvável	0	1
Pouco provável	0	0
Medianamente provável	0	0
Provável	1	0
Muito Provável	0,6	0
Altamente provável	0	0

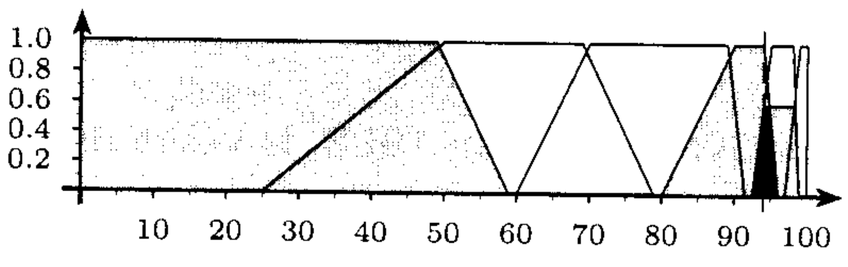

Figura 4: Fuzzificação para a variável *probabilidade*, no *Caso 2*.

Tabela 5: GPs para a variável *qualidade*, para o *Caso 2*.

	PQS
Ruim	0
Boa	0,67
Ótimo	1

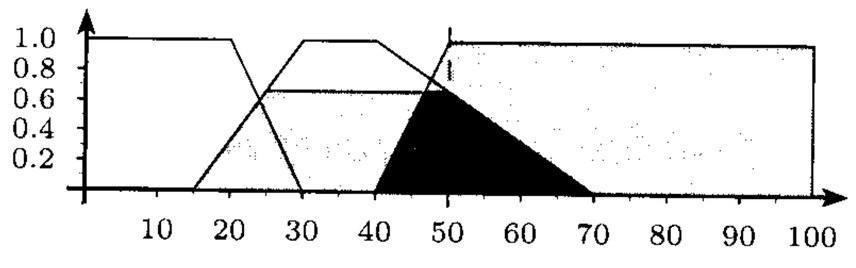

Figura 5: Fuzzificação para a variável *qualidade*, no *Caso 2*.

A *probabilidade*, para o *Caso 1*, é expressa pelos termos *muito provável* e *altamente provável*, e a *qualidade*, pelos termos *bom* e *ótimo* e, essas mesmas variáveis do *Caso 2*, pelos termos *improvável*, *provável*, *muito provável*, *bom* e *ótimo*.

4.2 INFERÊNCIA

A inferência executa operações sobre os conjuntos *fuzzy*, com a combinação dos antecedentes das regras, a implicação e a aplicação do *modus ponens* generalizado, sendo, esse procedimento, feito em dois passos: a "agregação", que corresponde ao operador lógico *E* que executa a intersecção entre conjuntos e, portanto, determina o mínimo entre os valores disparados pelas regras, seguido da "composição".

Os modelos ("máquinas") de inferência adequados para esse FIS, são os modelos de Mamdani ou de Larsen, visto que são sensíveis ao disparo de múltiplas regras sobre o conjunto de saída, quando, então, inicia-se o procedimento de defuzzificação, que começa com o segundo passo da inferência, a "composição", que é equivalente ao operador lógico *OU* e executa a união entre conjuntos, na qual o maior valor entre os mínimos resultantes da agregação é considerado para a defuzzificação.

Foram estabelecidas trinta e seis regras de inferência (Tabela 6), sendo que em metade dessas seus antecedentes são avaliados pelas variáveis *probabilidade* (PPS_1) e *qualidade* e, a outra metade, é avaliada pelas variáveis *probabilidade* (PPS_2) e *qualidade*. Essas regras, relacionam termos de entrada com a função de saída, expressa pelos termos *SNP descartado*, *SNP não confirmado* e *SNP confirmado*.

Tabela 6: Regras de inferência utilizadas no SID.

se	improvável	*e*	qualidade ruim	*então*	SNP descartado	(R_1)
	pouco provável		qualidade ruim		SNP descartado	(R_2)
	medianamente provável		qualidade ruim		SNP descartado	(R_3)
	provável		qualidade ruim		SNP descartado	(R_4)
	muito provável		qualidade ruim		SNP descartado	(R_5)
	altamente provável		qualidade ruim		SNP descartado	(R_6)
	improvável		qualidade boa		SNP descartado	(R_7)
	pouco provável		qualidade boa		SNP descartado	(R_8)
	medianamente provável		qualidade boa		SNP não confirmado	(R_9)
	provável		qualidade boa		SNP não confirmado	(R_{10})
	muito provável		qualidade boa		SNP confirmado	(R_{11})
	altamente provável		qualidade boa		SNP confirmado	(R_{12})
	improvável		qualidade ótima		SNP descartado	(R_{13})
	pouco provável		qualidade ótima		SNP descartado	(R_{14})
	medianamente provável		qualidade ótima		SNP não confirmado	(R_{15})
	provável		qualidade ótima		SNP não confirmado	(R_{16})
	muito provável		qualidade ótima		SNP confirmado	(R_{17})
	altamente provável		qualidade ótima		SNP confirmado	(R_{18})

No *Caso 1*, as funções de pertinência (Expressões 1, 2 e 3), resultam em $P_{MP} = 1$, para *PPS*$_2$, $P_{AP} = 1$, para *PPS*$_1$, $Q_B = 0,9$ e $Q_O = 0,3$ (Tabelas 2 e 3 e Figuras 2 e 3), então, a agregação é feita entre Q_B e Q_O, o que resulta no termo *ótima* para a variável *qualidade*. Os demais valores obtidos são iguais e, assim, não aplica-se a agregação, o que resulta em *muito provável* (*PPS*$_2$) e *altamente provável* (*PPS*$_1$), para *probabilidade*, que disparam as regras R_{17} e R_{18}.

Para o *Caso 2*, após a agregação, toma-se $P_{IM} = 1$ (*PPS*$_2$), $P_{MP} = 0,6$ (*PPS*$_1$) e $Q_B = 0,67$ que são levados à máquina de inferência, que dispara R_7 e R_{11}.

O modelo de inferência mapeia os antecedentes, resultantes da agregação, no termo conseqüente, que, para os modelos de Mamdani ou Larsen, representa uma função de saída em termos lingüísticos, exatamente como uma função de pertinência.

A função de saída que foi estabelecida, reduz as seis classes definidas para o *PPS* aos termos *SNP descartado* (SNP_D), *SNP não confirmado* (SNP_{NC}) e *SNP confirmado* (SNP_C), que, então, compõem a variável lingüística *SNP* (Expressões 4):

$$SNP_D(x) = \begin{cases} 1 & x \leq 20 \\ \frac{30-x}{30-20} & 20 < x < 30 \\ 0 & x \geq 30 \end{cases} \quad SNP_{NC}(x) = \begin{cases} 0 & x \leq 15 \\ \frac{x-15}{30-15} & 15 < x < 30 \\ 1 & 30 \leq x \leq 40 \\ \frac{70-x}{70-40} & 40 < x < 70 \\ 0 & x \geq 70 \end{cases} \quad SNP_C(x) = \begin{cases} 0 & x \leq 40 \\ \frac{x-40}{50-40} & 40 < x < 50 \\ 1 & x \geq 50 \end{cases} \tag{4}$$

As regras R_{17} e R_{18}, disparadas no *Caso 1*, são processadas como:

1. R_{17} tem como antecedentes o valor *muito provável*, com $GP = 1$, e o valor *ótima*, com $GP = 0,3$; assim, a aplicação da regra mapeia o conseqüente *SNP confirmado*, com $GP = 1$ e $GP = 0,3$, isto é $SNP_C = 1$ e $SNP_C = 0,3$;

2. R_{18} tem como antecedentes o valor *altamente provável*, com $GP = 1$, e o valor *ótima*, com $GP = 0,3$; então, da mesma forma, mapeia o conseqüente *SNP confirmado*, com $GP = 1$ e $GP = 0,3$, isto é $SNP_C = 1$ e $SNP_C = 0,3$.

Com a aplicação das duas regras, cujos resultados foram coincidentes, apenas o termo *SNP confirmado* foi mapeado e o procedimento de composição deve ser tomado somente sobre esse termo. A composição busca o máximo entre os GPs de cada termo, no caso, somente sobre o termo *SNP confirmado*, fazendo $SNP_C = 1$.

Para o *Caso 2*, são disparadas as regras R_7 e R_{11}, que avaliam os valores antecedentes $P_{IM} = 1$ e $Q_O = 0,67$, para R_7, e $P_{MP} = 0,6$ e $Q_O = 0,67$, para R_{11}. A regra R_7 mapeia na função de saída o valor *SNP descartado*, com $GP = 1$ e $GP = 0,67$, enquanto a regra R_{11} mapeia na função de saída o valor *SNP não confirmado*, com $GP = 0,67$ e $GP = 0,6$. O termo *SNP confirmado* não foi mapeado, logo o procedimento de composição aplicado aos demais termos resulta em *SNP descartado*, com $GP = 1$ ($SNP_D = 1$), e *SNP não confirmado*, com $GP = 0,67$ ($SNP_{NC} = 0,67$).

As Figuras 6 e 7 representam, respectivamente, a aplicação das regras de inferência sobre a função de saída (Expressões 4) para os *Casos 1* e *2*.

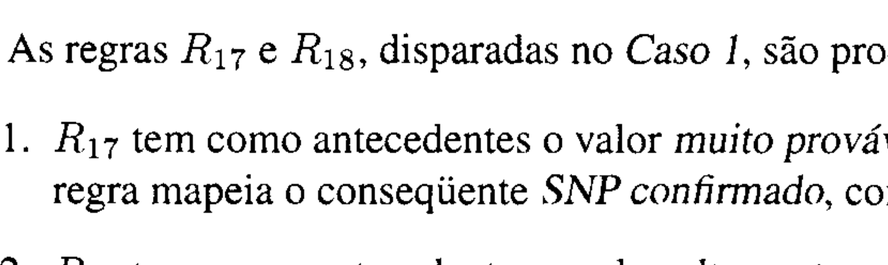

Figura 6: Aplicação das regras de inferência para o *Caso 1*.

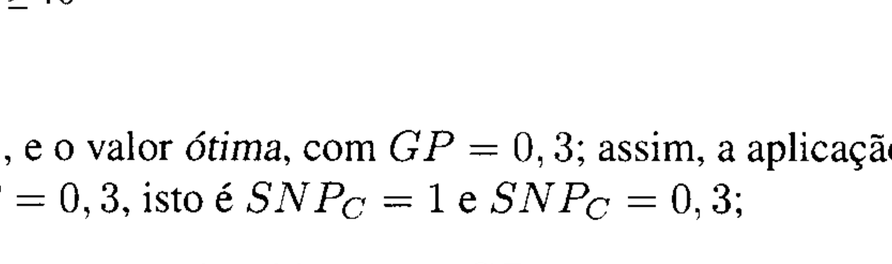

Figura 7: Aplicação das regras de inferência para o *Caso 2*.

4.3 DEFUZZIFICAÇÃO

A defuzzificação executa a composição, que determina os valores que representam cada um dos conjuntos mapeados na função de saída, e, a partir desses, calcula um valor preciso (VP), obtido com a aplicação do método de defuzzificação.

Para o modelo proposto, o método de defuzzificação deve considerar múltiplos disparos, pois o valor da qualidade da base no consenso é utilizada como um "valorizador" dos valores de probabilidade confrontados (PPS_1 e PPS_2). Assim, havendo disparos múltiplos, esses devem ser avaliados, pois, servem à idéia de valorizar os conjuntos difusos estabelecidos na função de saída. Para esse fim, deve ser utilizado o método centro de máximo (*center of maximum* - COM) e, a partir dos modelos de inferência, aplica-se o método de defuzzificação adequado ao problema. Como o fuzzyMorphic.pl permite a inferência pelos modelos de Mamdani e Larsen, ambos podem ser aplicados e, juntamente com os valores tomados da composição, definem os valores para a defuzzificação.

O COM (Expressão 5), trata-se de uma média ponderada, onde o numerador é o somatório dos valores da composição (h_i), isto é, a altura do conjuntos de saída, multiplicados pelos valores no universo de discurso (u_i), encontrados pelo modelo de inferência, do seu respectivo conjunto de saída, e o denominador é o somatório das alturas (h_i).

Para o *Caso 1*, o VP (Expressão 6) e sua representação (Figuras 8) são iguais para os modelos de Mamdani e Larsen, mas, para o *Caso 2*, como conseqüência desses modelos, a defuzzificação apresenta diferentes resultados (Expressões 7 e 8 e Figuras 9 e 10).

$$VP = \frac{\sum h_i \cdot u_i}{\sum h_i} \quad (5) \qquad VP_{C_1} = \frac{75 \cdot 1}{1} = 75 \quad (6)$$

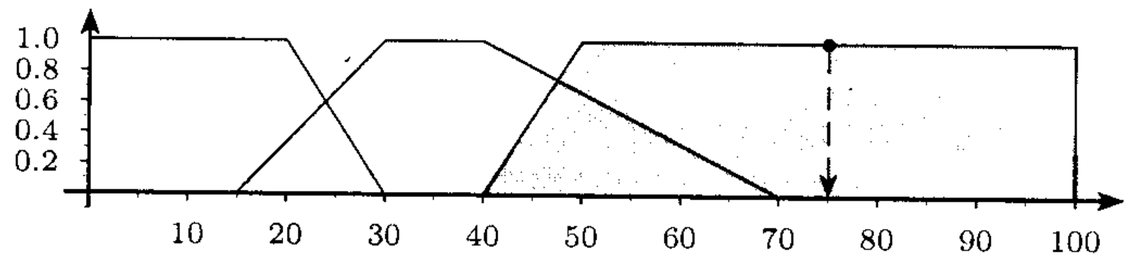

Figura 8: Aplicação do modelo de inferência, para o *Caso 1*.

$$VP_{C_2} = \frac{(10 \cdot 1) + (37,475 \cdot 0,67)}{1 + 0,67} = 21,02 \quad (7) \qquad VP_{C_2} = \frac{(10 \cdot 1) + (35 \cdot 0,67)}{1 + 0,67} = 20,03 \quad (8)$$

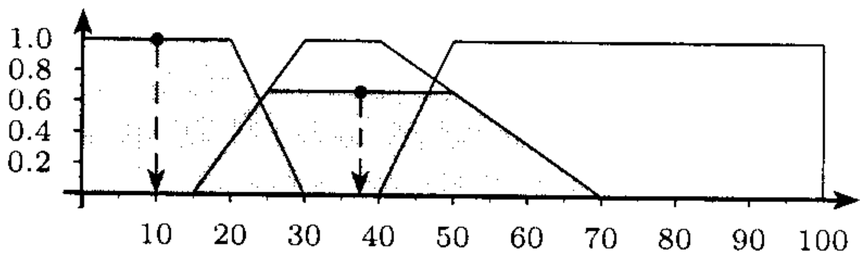

Figura 9: Aplicação do modelo de Mamdani para o *Caso 2*.

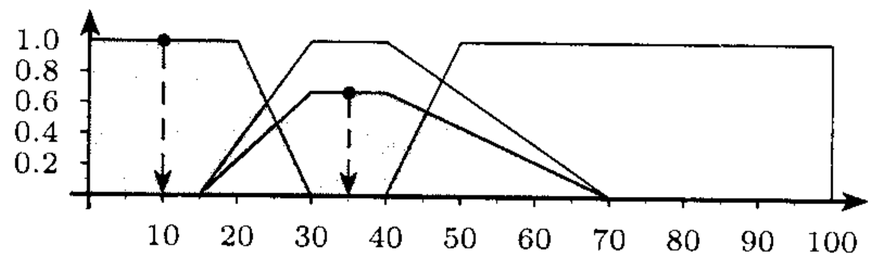

Figura 10: Aplicação do modelo de Larsen para o *Caso 2*.

5 DISCUSSÃO SOBRE O MODELO E O SISTEMA *FUZZY* PARA IDENTIFICAÇÃO DE SNPs

O *Caso 1*, apresenta uma situação na qual as investigações prévias chegam a resultados próximos e as probabilidades da posição vir a ser um SNP são de 99% e 96%. Entretanto, o *Caso 2*, é bastante discordante, pois, os resultados são 94% e zero.

O SID incluiu um novo atributo ao objeto, agregando outra informação e, portanto, amplia as possibilidades de investigação. A *PQS*, que informa a qualidade da base no consenso, é utilizada como um atributo "valorizador" na análise a ser feita como apoio à decisão, assumindo $PQS = 43$ para o *Caso 1* e $PQS = 50$ para o *Caso 2*. Assim, aos resultados prévios da possibilidade de o ponto vir a ser um SNP, acrescenta-se a qualidade do ponto em questão, analisando-a com o objetivo de informar uma das três possibilidades excludentes: a confirmação do SNP, a eliminação dessa possibilidade ou, uma situação intermediária, sem a confirmação dessa possibilidade, mas também sem elementos conclusivos para seu descarte.

A análise dos *Caso 1* e *2*, fornece elementos suficientes para demonstrar o uso do modelo implementado, contudo, devem ser esperados resultados efetivos de seu uso, a partir da análise de conjuntos de dados, quando as decisões tomadas pelo modelo de inferência podem, então, ser agrupadas, determinando os conjuntos de pontos que melhor se ajustam as três possibilidades procuradas [10].

Estabelecer agrupamentos é uma tarefa complexa, pois procura-se dizer como são e em quantas classes os dados se distribuem, sem que se tenha conhecimento a priori dos mesmos. Além disso, as classes podem não existir, caso os elementos se distribuam eqüitativamente por todo o espaço, não caracterizando qualquer categoria, pois as classes são construídas com base na semelhança entre os elementos, cabendo a verificação das possíveis classes resultantes para avaliar a existência de algum significado útil [14].

6 CONCLUSÕES

Critérios fixos e precisos de classificação, em geral, não são adequados, quando um estudo apresenta resultados próximos à divisão das classes, o que pode ser tratado por SIDs, que também são convenientes e possuem capacidade para tratar problemas que apresentam incerteza ou imprecisão para a tomada de decisão.

Ao adicionar um novo atributo aos resultados prévios, o sistema *fuzzy* é capaz de decidir, de forma única, entre as três possibilidades resultantes do modelo e, então, agrupá-las a partir de um algoritmo não-supervisionado e com estabelecimento dinâmico do número de grupos, esperando que o resultado desse agrupamento confirme o particionamento do conjunto em três grupos, não necessitando de limites fixos e precisos para a identificar possíveis SNPs.

REFERÊNCIAS

[1] M. C. Naldi and A. P. L. F. de Carvalho. "Utilização de algoritmos de aprendizado de máquina evolutivos para análise de nível de expressão gênica". In *Anais...*, São Leopoldo, 2005. XXV Congresso da Sociedade Brasileira de Computação, Universidade do Vale dos Sinos.

[2] W. Arbex. "fuzzyMorphic.pl". 1 CD, 2009. Perl. Ambiente UNIX-like com GUI e interpretador Perl 5.0 ou posterior.

[3] P. E. M. de Almeida and A. G. Evsukoff. *Sistemas fuzzy*, pp. 169–202. Manole, Barueri, 2005.

[4] A. J. Brookes. "The essence of SNPs". *Gene*, vol. 2, no. 234, pp. 177–186, July 1999.

[5] T. Brown. *Genomes*. John Wiley & Sons, New York, second edition, 2002.

[6] L. A. Zadeh. "Outline of a new approach to the analysis of complex systems and decision processes". *IEEE Trans. on Systems, Man, and Cybernetics*, vol. SMC-3, pp. 28–44, 1973.

[7] G. J. Klir and B. Yuan. *Fuzzy sets and fuzzy logic: theory and applications*. Prentice Hall, Upper Saddle River, May 1995.

[8] D. A. Nickerson, S. L. Taylor, N. Kolker, J. Sloan, T. Bhangale, M. Stephens and I. Robertson. *Polyphred users manual*. University of Washington, Seattle, May 2008. Version 6.15 Beta.

[9] R. Tanscheit. *Sistemas fuzzy*, pp. 229–264. Thomson Learning, São Paulo, 2007.

[10] W. Arbex. "Modelos computacionais para identificação de informação genômica associada à resistência ao carrapato bovino". Doutorado em Engenharia de Sistemas e Computação, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Mar. 2009. Trabalho vencedor do prêmio SBIAgro 2009 - Categoria Tese de Doutorado no VII Congresso da Sociedade Brasileira de Agroinformática.

[11] D. A. Nickerson, V. O. Tobe and S. L. Taylor. "PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing". *Nucl. Acids Res.*, vol. 25, no. 14, pp. 2745–2751, 1997.

[12] G. T. Marth, I. Korf, M. D. Yandell, R. T. Yeh, Z. Gu, H. Zakeri, N. O. Stitziel, L. Hillier, P.-Y. Kwok and W. R. Gish. "A general approach to single-nucleotide polymorphism discovery". *Nature Genetics*, vol. 23, pp. 452–456, Dec. 1999.

[13] B. Ewing, L. Hillier, M. C. Wendl and P. Green. "Base-calling of automated sequencer traces using Phred (I): accuracy assessment". *Genome Research*, vol. 8, no. 3, pp. 175–185, March 1998.

[14] L. A. V. de Carvalho. *Datamining: a mineração de dados no marketing, medicina, economia, engenharia e administração*. Ciência Moderna, Rio de Janeiro, 2005.