Embrapa

Gado de Leite

Boletim de Pesquisa e Desenvolvimento 29
ISSN 0104-9046
Novembro, 2009

Inferência difusa aplicada à identificação de informação genômica em sequências de cDNA

Ministério da Agricultura, Pecuária e Abastecimento

BRASIL UM PAÍS DE TODOS GOVERNO FEDERAL

Embrapa

ISSN 0104-9046
Novembro, 2009

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Gado de Leite
Ministério da Agricultura, Pecuária e Abastecimento

Boletim de Pesquisa e Desenvolvimento 29

Inferência difusa aplicada à identificação de informação genômica em sequências de cDNA

Wagner Antonio Arbex
Leonardo Gerheim de Andrade
Ricardo Ferreira Tagliatti
Marcos Vinícius Gualberto Barbosa da Silva
Marta Fonseca Martins Guimarães
Luís Alfredo Vidal de Carvalho

Embrapa Gado de Leite
Juiz de Fora, MG
2009

Exemplares desta publicação podem ser adquiridos na:
Embrapa Gado de Leite
Rua Eugênio do Nascimento, 610 – Bairro Dom Bosco
36038-330 Juiz de Fora, MG
Fone: (32) 3311-7400
Fax: (32) 3311-7401
Home page: http://www.cnpgl.embrapa.br
E-mail: sac@cnpgl.embrapa.br

Supervisão editorial: Wagner Arbex
Normalização bibliográfica: Inês Maria Rodrigues
Editoração eletrônica: Carlos Alberto Medeiros de Moura
Capa: Moema Sarrapio Pereira
1ª edição
1ª impressão (2009): 1.000 exemplares

Dados Internacionais de Catalogação na Publicação (CIP)
Embrapa Gado de Leite

Inferência difusa aplicada à identificação de informação genômica em sequências de cDNA / Wagner Antonio Arbex... [et al.]. – Juiz de Fora : Embrapa Gado de Leite, 2009.
30 p. (Embrapa Gado de Leite. Boletim de Pesquisa e Desenvolvimento, 29).

ISSN 0104-9046

1. Sistema de inferência difusa. 2. Mineração de dados. 3. Suporte à decisão. 4. Polimorfismo de base única. 5. Bioinformática. I. Wagner Antonio Arbex. II. Leonardo Gerheim de Andrade. III. Ricardo Ferreira Tagliatti. IV. Marcos Vinícius Gualberto Barbosa da Silva. V. Marta Fonseca Martins Guimarães. VI. Luiz Alfredo Vidal de Carvalho. VII. Série.

CDD 570.285

Sumário

Inferência difusa aplicada à identificação de informação genômica em sequências de cDNA

Wagner Antonio Arbex[1]
Leonardo Gerheim de Andrade[2]
Ricardo Ferreira Tagliatti[2]
Marcos Vinícius Gualberto Barbosa da Silva[3]
Marta Fonseca Martins Guimarães[4]
Luís Alfredo Vidal de Carvalho[5]

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 - SNP*), são importantes no estudo da variabilidade das espécies, pois podem provocar alterações funcionais ou fenotípicas, que, por sua vez, podem implicar em consequências evolutivas ou bioquímicas nos indivíduos das espécies. A descoberta de SNPs por algoritmos computacionais é uma prática bastante difundida e as implementações desses algoritmos trabalham com diferentes metodologias, 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, quando os resultados são discordantes. Os resultados apresentados são originários de um modelo computacional de aprendizado de máquina que implementa uma metodologia de mineração

[1] Matemático, D.Sc – Analista da Embrapa Gado de Leite – arbex@cnpgl.embrapa.br
[2] Estudante de Análise de Sistemas – Estagiário do Laboratório de Bioinformática e Genômica Animal da Embrapa Gado de Leite – bioinf@cnpgl.embrapa.br
[3] Zootecnista, D.Sc – Pesquisador da Embrapa Gado de Leite – marcos@cnpgl.embrapa.br
[4] Bióloga, D.Sc – Pesquisadora da Embrapa Gado de Leite – mmartins@cnpgl.embrapa.br
[5] Engenheiro Mecânico, D.Sc – Professor da Universidade Federal do Rio de Janeiro – alfredo@cos.ufrj.br

de dados, fundamentado em lógica difusa (*fuzzy logic*), para auxiliar na tomada de decisão, no caso em que as informações sejam divergentes e, também, na confirmação de informações coincidentes.

Palavras-chave: sistema de inferência difusa, mineração de dados, suporte à decisão, polimorfismo de base única, bioinformática.

Introdução

Caracterização do problema

Em bovinos, existe uma variabilidade genética que, se adequadamente identificada, pode contribuir para a sustentabilidade dos sistemas de produção animal, pois pode tornar o animal resistente a parasitas, reduzindo ou eliminando o uso de produtos químicos e medicamentos para evitar ou tratar a infestação.

A variação genética existente entre as raças de *Bos taurus* e *Bos indicus* para as características associadas à resistência a esses parasitas e as atuais ferramentas da genética molecular sugerem a utilização de marcadores genéticos, associados à resistência, como auxílio aos programas de melhoramento, visando à obtenção de animais resistentes e, consequentemente, economicamente mais produtivos (MARTINEZ et al., 2004; TEODORO et al., 2004).

Em virtude do exposto, essa pesquisa[1] propôs e desenvolveu um modelo computacional capaz de investigar informações genômicas relacionadas a genes candidatos, cuja expressão esteja associada à resistência ao carrapato *Boophilus microplus*, o que auxiliará a seleção de animais na tentativa de se obterem indivíduos mais adaptados às condições desfavoráveis que são tipicamente encontradas nos trópicos.

O modelo computacional desenvolvido utiliza-se de lógica difusa como base para a implementação de um sistema de inferência, auxiliar à

1 Uma descrição completa de todas as ações e atividades dessa pesquisa pode ser vista em Arbex (2009).

tomada de decisão e que faz mineração de dados, por meio de busca não-supervisionada, partindo de consensos de sequências expressas identificadas (*expressed sequence tags – ESTs*), originadas de cDNA, de animais susceptíveis ou de animais resistentes, que serão contrastadas em busca de SNPs na tentativa de identificar os genes que se expressaram e que, portanto, conferem resistência ou susceptibilidade aos animais frente a uma infestação por carrapatos.

Para que fosse possível alcançar essa proposta, cerca de 400 animais F_2, provenientes do cruzamento de animais Holandês (*Bos taurus*) com animais Gir (*Bos indicus*), foram avaliados quanto à resistência a carrapatos, quando submetidos a infestação artificial. Após a avaliação dos mesmos, foi coletado material biológico, de onde foi extraído o RNA dos animais que apresentaram extremos de comportamento em relação à infestação, ou seja, maior susceptibilidade e maior resistência. A partir do RNA extraído, em particular, do mRNA, é feita a transcrição reversa para a obtenção do cDNA que, então, deve ser utilizado para a geração das ESTs.

Motivação

No Brasil, o controle dos carrapatos por meio da seleção de animais e raças resistentes ocorre em menor grau em relação ao uso de produtos químicos (MARTINEZ et al., 2004), no caso específico, de acaricidas. Por conseguinte, o uso contínuo dos acaricidas tem provocado o aparecimento de carrapatos resistentes ao princípio químico ativo dos mesmos – em geral, causam o efeito esperado por pouco mais de uma década (VAZ et al., 2000) – sendo necessário o desenvolvimento de novos produtos. Assim, a cada novo produto, novas cepas resistentes aparecem, perpetuando esse ciclo vicioso (MARTINEZ et al., 2004; ATHAYDE et al., 2001; VAZ et al., 2000).

A aplicação de substâncias químicas acaricidas, tem sido largamente utilizado no combate dos mesmos. Todavia, além de adquirir resistência ao princípio ativo, a eficácia de tais substâncias é questionada já que não conseguem eliminar os parasitas por completo (MARTINEZ et al., 2004; ATHAYDE et al., 2001; VAZ et al., 2000) e que, o uso prolon-

gado desses produtos, além de representar um custo considerável, uma vez que, no país, o gasto anual com produtos químicos para o combate aos parasitas é da ordem de R$ 800 milhões (MARTINEZ et al., 2004), pode causar efeitos colaterais, já que deixam resíduos químicos que podem vir a contaminar a carne, o leite e o meio ambiente (MARTINEZ et al, 2004; VAZ et al., 2000).

Assim, a necessidade de que sejam criadas alternativas para o controle dos carrapatos é notória (MARTINEZ et al., 2004), o que confere importância à iniciativa da Embrapa Gado de Leite, quando, em 1995, iniciou um projeto com várias ações de pesquisa, visando à identificação de marcadores genéticos em bovinos associados à resistência a endoparasitas, ectoparasitas e ao estresse térmico, sendo a pesquisa ora apresentada uma dessas ações.

Objetivos

Um dos principais resultados que normalmente se pretende alcançar, especificamente a partir de estudos em genômica realizados nos mamíferos, é caracterizar o padrão de expressão gênica que corresponde a eventos fisiológicos importantes relacionados à produção e à saúde dos animais.

Em consonância com esse propósito, o objetivo primário do presente projeto de pesquisa é o desenvolvimento e a implementação de um modelo computacional aplicado à investigação de polimorfismos de base única em sequências expressas de cDNA, frente à ação do carrapato bovino.

Como consequência direta, os resultados obtidos a partir do modelo podem vir a permitir a identificação de informações genômicas que estejam associadas à resistência ao carrapato bovino e que possibilitem a "marcação" de sequências genéticas ou de genes relacionados com a proposta em questão e, servindo ainda para auxiliar projetos que busquem o desenvolvimento de mecanismos alternativos de controle e eliminação do carrapato *Boophilus microplus* em bovinos. Tais projetos,

por sua vez, poderão estabelecer novos e factíveis meios economicamente viáveis e auto-sustentáveis para o controle da infestação do *Boophilus microplus* em bovinos.

Modelo de inferência difusa para identificação de candidatos a SNPs

Procedimento metodológico do sistema de inferência difusa (SID) para mineração de dados

A estrutura funcional do modelo de mineração de dados está representada nas Figs. 1 e 2, nas quais pode-se destacar a divisão de seu funcionamento em etapas bem definidas, quais sejam:

a) o processamento inicial dos cromatogramas, quando é feita a leitura das bases e, consequentemente, são originadas as sequências, as sequências-consenso (*contigs*) e, ainda, determina a qualidade das bases dessas sequências. Essa etapa é feita pelo *pipeline phredPhrap* e são gerados diversos arquivos, entre eles, o arquivo formato *ace* e os diversos arquivos *phd*, um para cada sequência lida (Fig. 1);

b) a execução dos programas *Polyphred* e *Polybayes* sobre os arquivos *ace* e *phd*, e cada um desses programas, de acordo com a sua metodologia, identifica os pontos candidatos a SNPs e estabelece uma probabilidade para cada um desses pontos. Esses resultados são registrados nos arquivos *polyphred.out* e *report.out*, que serão utilizados como dados de entrada para o procedimento de mineração de dados (Fig. 1);

c) na etapa seguinte é feita a preparação dos dados, quando os dados oriundos do *Phrap* (obtidos na execução do *pipeline phredPhrap*), do *Polyphred* e do *Polybayes* são extraídos e selecionados dos seus respectivos arquivos e, ainda, se necessário, complementados. Essa etapa de preparação dos dados é feita pelos *scripts parsepolyBayes.pl, parsepolyPhred.pl, parsephrapQality.pl* e *joinparsersOut.pl*, que, ainda, estrutura o arquivo para que seja lido pelo *fuzzyMorphic.pl* (Fig. 2);

d) no passo seguinte, com a execução do *fuzzyMorphic.pl*, é feito o procedimento de mineração de dados, implementado em um SID, que fornece como saída um arquivo com os mesmos dados de entrada,

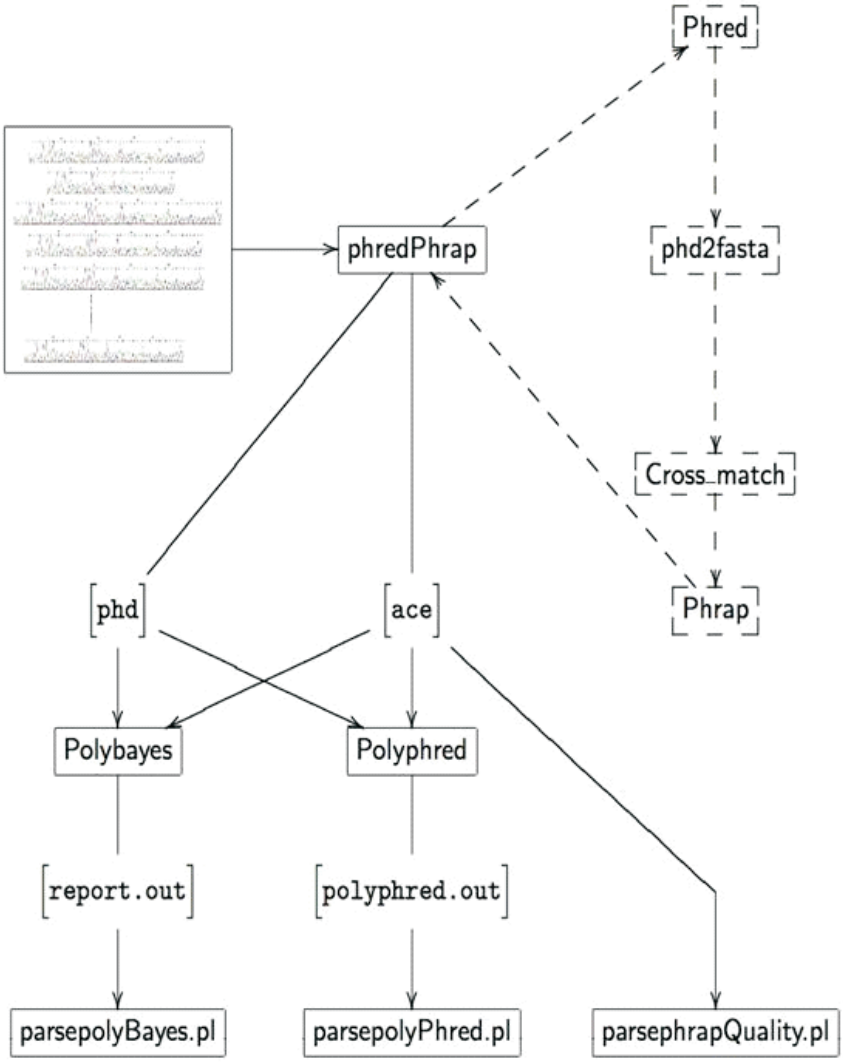

Fig. 1. Síntese da estrutura funcional do modelo de mineracão de dados (etapas A e B)

e) a última etapa de análise e avaliação dos resultados não integra o estudo desse projeto e utiliza técnicas e ferramentas para verificação dos resultados inferidos conhecidos e com eficácia comprovada. No

caso, é feita uma análise de agrupamento sobre o conjunto de dados resultante do processamento do sistema de inferência (Fig. 2).

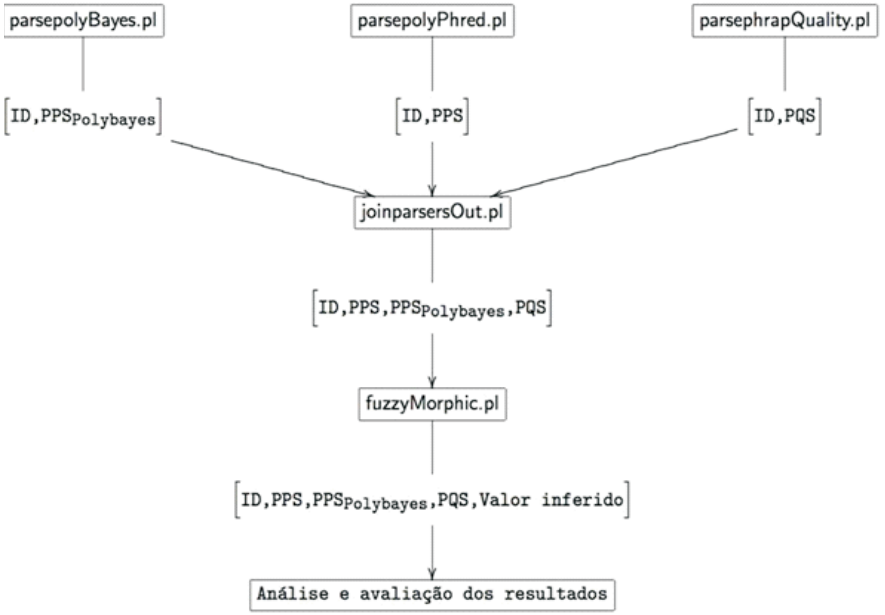

Fig. 2. Síntese da estrutura funcional do modelo de mineracão de dados (etapas C, D e E).

Aplicação do SID para descoberta de candidatos a SNP

Os cromatogramas processados são, ao todo, 4.072, sendo 1.884 cromatogramas oriundos de animais que apresentaram maior resistência ao carrapato bovino e 2.188 cromatogramas oriundos de animais que apresentaram menor resistência ao carrapato bovino.

No *pipeline phredPhrap* deve ser estabelecido um escore mínimo, cujo parâmetro é conhecido como *minscore*, para a região alinhada, o que interfere diretamente no número de sequências agrupadas no alinhamento e, como consequência direta, interfere na profundidade da cobertura. Quanto maior for o escore estabelecido, mais rigoroso será

o alinhamento, criando uma tendência de se eliminarem sequências que não coincidem com outras e, portanto, não são alinhadas e não participam da montagem das sequências-consenso.

Por esse fato, a título de comparação, o modelo de inferência foi aplicado sobre três diferentes conjuntos de alinhamentos gerados pelo *phredPhrap* com escore mínimo de 30%, que é o padrão do programa, e outros dois com escores mínimos de 60% e 90%.

Com a execução do *pipeline*, com o *minscore* em 30% para o *minscore*, foram obtidas 502 sequências-consenso, a partir de 1.810 sequências, pois 2.262 sequências, cerca 55%, foram eliminadas por não atenderem ao escore mínimo para alinhamento ou, eventualmente, por algum problema de sequenciamento que tenha sido detectado. A Tabela 1 relaciona a quantidade de sequências utilizadas no alinhamento, com a quantidade de sequências-consenso obtidas, isto é, o número de *contigs* montados. Por exemplo, a segunda linha da Tabela 1 informa que foram montadas 261 sequências-consenso e, para cada uma dessas foram utilizadas duas sequências.

A partir das sequências alinhadas, foram executados os programas *Polyphred* e *Polybayes*, que, sem restrições, permitiram que todos os pontos das sequências fossem considerados possíveis SNPs, independentemente da probabilidade aferida por esses programas. Em geral, para a execução desses dois programas, determina-se um parâmetro de corte de 70% e, assim, os pontos cuja probabilidade for inferior a esse piso são automaticamente descartados e não são considerados SNPs, sob nenhum aspecto. Entretanto, como a ideia é avaliar o modelo implementado em situações adversas, não foi estipulado nenhum parâmetro de corte e o sistema de inferência deve avaliar, inclusive, pontos que apresentam zero de probabilidade de vir a ser um SNP. A única restrição estabelecida ocorre quando ambos coincidem e determinam zero de probabilidade para o mesmo ponto e, nesse caso, os pontos são eliminados na etapa de preparação dos dados.

Tabela 1. Número de sequências-consenso obtidas, em relação ao número de sequências agrupadas no seu alinhamento, obtidas com escore mínimo de alinhamento em 30%.

Número de sequências alinhadas	Número de sequências-consenso obtidas
1	32
2	261
3	80
4	44
5	25
6	19
7	5
8	4
9	5
10	3
11	4
12	4
14	4
15	4
16	1
17	2
18	2
20	1
21	2
22	1
24	1
25	1
30	1
82	1

Assim, considerando as sequências alinhadas com o *minscore* em 30%, o *Polyphred* identificou 1.503 candidatos a SNPs, com o *PPS*[2] entre 1% e 99%, e o *Polybayes*, identificou 240 candidatos, com o $PPS_{Polybayes}$ entre 2% e 100%. Quando esses resultados foram combinados, durante a preparação dos dados para o procedimento de mineração, foi gerado um conjunto contendo 1.624 possíveis pontos polimórficos, que, então, deve ser submetido ao modelo de inferência, implementado pelo *fuzzyMorphic.pl.* Os respectivos conjuntos de candidatos a SNPs retornados pelos dois programas podem apresentar intersecção, total ou parcial, ou, ainda, não se interceptarem. Nesse caso, a observação do número de elementos resultante da combinação

[2]O *Polyphred Score* (*PPS*) (NICKERSON et al., 2008) é um valor estabelecido pelo *Polyphred*, entretanto, o *Polybayes* possui uma medida para o mesmo fim que será denominada de $PPS_{Polybayes}$ para simplificar a terminologia em uso

dos conjuntos estabelecidos pelo *Polyphred* e pelo *Polybayes*, mostra que esses conjuntos se interceptam parcialmente.

As Tabelas 2 e 3, sintetizam resultados e quantidades apresentados nesta seção referentes aos conjuntos de dados obtidos na etapa de preparação para os três valores adotados para o *minscore* e que, assim, podem ser submetidos ao procedimento de mineração de dados.

Tabela 2. Sequências-consenso obtidas para mineração dos SNPs.

Escore mínimo de alinhamento	Sequências consideradas	Sequências descartadas	Consensos
30%	1.810	2.262	502
60%	1.649	2.423	470
90%	1.547	2.525	447

Tabela 3. Candidatos a SNPs.

Escore mínimo de alinhamento	Candidatos a SNP				Intervalo de variação	
	Polyphred	Polybayes	Total	Polyphred ∧ Polybayes	PPS (%)	$PPS_{Polybayes}$ (%)
30%	1.503	240	1.624	119	1 – 99	2 – 100
60%	1.224	230	1.378	76	1 – 99	2 – 100
90%	1.055	202	1.145	112	1 – 99	3 – 100

Da mesma forma como apresentado na Tabela 1, considerando o *minscore* em 30%, as Tabelas 4 e 5 relacionam a quantidade de sequências utilizadas no alinhamento, com a quantidade de sequências-consenso obtida, para o *minscore* em 60% e 90%, respectivamente.

Tabela 4. Número de sequências-consenso obtidas, em relação ao número de sequências agrupadas no seu alinhamento, obtidas com escore mínimo de alinhamento em 60%.

Número de sequências alinhadas	Número de sequências-consenso obtidas
1	22
2	260
3	75
4	41
5	18
6	15
7	2
8	7
9	5
10	4
11	3
12	1
13	2
14	4
16	2
17	3
19	2
21	2
26	1
30	1
54	1

Tabela 5. Número de sequências-consenso obtidas, em relação ao número de sequências agrupadas no seu alinhamento, obtidas com escore mínimo de alinhamento em 90%.

Número de sequências alinhadas	Número de sequências-consenso obtidas
1	11
2	258
3	68
4	37
5	20
6	14
7	4
8	6
9	3
10	5
12	3
13	3
14	3
15	2
16	4
17	2
19	1
21	1
26	1
30	1

Análise e avaliação dos resultados obtidos

O modelo de aprendizado de máquina implementado, sob o aspecto funcional, investiga o conjunto de dados originado a partir da junção dos conjuntos gerados pelo *Polyphred* e pelo *Polybayes*; avalia as probabilidades, estabelecidas por suas diferentes propostas, de cada elemento do conjunto; e, então, determina para cada um dos elementos um novo atributo que deve servir como uma referência na tentativa de particionar o conjunto de dados em grupos de elementos que podem ser tratados como pontos polimórficos confirmados (*SNP confirmado*), pontos não polimórficos (*SNP descartado*) e, ainda, pontos sem elementos suficientes para uma definição conclusiva (*SNP não confirmado*).

Entretanto, qualquer classificação que se queira fazer pode vir a ser influenciada pela forma ou comportamento dos dados ou, ainda, para classes definidas com limites precisos, promover decisões duvidosas, quando o valor estiver muito próximo dos limites das classes. Tais questões, entre outras, sugerem a adoção de métodos de particionamento não-hierárquicos e não-supervisionados, pois não partem de nenhuma premissa externa para estabelecer as classes que podem particionar um conjunto, mas, de forma oposta, suas premissas são estabelecidas por características específicas, internas e inerentes ao conjunto avaliado, eliminando ou reduzindo a ação de agentes externos ao modelo, como a definição *a priori* de limites precisos para as classes.

As premissas dos métodos de particionamento decorrentes de algoritmos não-hierárquicos baseiam-se no próprio conjunto de valores avaliados, buscando a máxima coesão interna dos objetos de um grupo e o máximo isolamento entre os grupos (CARVALHO, 2005). Em outras palavras, a partir da análise do próprio conjunto, busca-se identificar os elementos que, em relação ao atributo avaliado, possuem a "menor distância" entre os elementos do grupo e, uma vez estabelecidos grupos cujos elementos possuem essa característica, tais grupos devem apresentar a "maior distância" entre si. Assim, como essas premissas

são decorrentes dos próprios valores analisados, reduzem o efeito de comportamento dos dados, isto é, supondo que o atributo avaliado apresente uma determinada tendência, todos os os elementos possuem o mesmo comportamento e um particionamento não direcionado e tomado a partir dos próprios elementos pode reduzir ou eliminar essa tendência.

A exclusão de premissas externas pode ser benéfica, assim como os modelos adotados para a avaliação de resultados devem ser reduzidos, no sentido de que devem ser simples e envolver poucos parâmetros. Dentro do possível, devem ser auto-contidos, não dependendo de componentes externos e utilizarem o menor número de variáveis e parâmetros possíveis, evitando que sejam criadas "condições de contorno" que permitam "acomodar" um resultado, ao invés de, verdadeiramente, encontrá-lo.

Estabelecer agrupamentos de dados é uma tarefa complexa e de difícil implementação, pois procura-se dizer como são e em quantas classes os dados se distribuem, sem que se tenha conhecimento prévio dos mesmos. As classes podem, simplesmente, não existir, caso os elementos se distribuam equitativamente por todo o espaço e não caracterizem qualquer categoria, pois os grupos ou classes são construídos com base na semelhança entre os elementos, cabendo, posteriormente, a verificação das possíveis classes resultantes para avaliar a existência de algum significado útil (CARVALHO, 2005).

Sob essa análise, o modelo de mineração de dados, implementado a partir de técnicas de aprendizado de máquina, substitui, através de inferência difusa, uma medida de probabilidade, contínua no intervalo [0,1] e associada à possibilidade de um ponto vir a ser um SNP, por um outro atributo, que permite agrupar os pontos dentro de três partições: *SNP confirmado*, *SNP descartado* e *SNP não confirmado*. Portanto, após o processamento dos dados pelo SID, o que se propôs foi executar o agrupamento dos dados resultantes por um algoritmo não-supervisionado e com estabelecimento dinâmico do número de grupos, esperando que o resultado obtido confirme o particionamento do conjunto em três grupos, baseado no novo atributo.

A operacionalização desse procedimento é feita por meio do *fuzzyMorphic.pl*, que implementa o SID e estabelece esse novo atributo. Já a análise do agrupamento é feita com o auxílio do *Weka* (*Waikato Environment for Knowledge Analysis*) (WITTEN; FRANK, 2005), versão 3.4.14, desenvolvido na Universidade de Waikato e distribuído sob *GNU General Public License*.

Entre os algoritmos para agrupamento, o *Weka* implementa o algoritmo *Expectation-Maximization (EM)*, que possui a propriedade de estabelecer, durante a execução, o número de grupos que melhor acomoda os elementos analisados, sem que essa informação seja fornecida. O algoritmo *EM* foi desenvolvido para problemas gerais de inferência estatística e seu funcionamento busca localizar o valor de um parâmetro que maximize a função de verossimilhança. Para o procedimento de agrupamento, foi adotado o padrão de dividir o conjunto de dados em 2/3 para treinamento e 1/3 para teste.

As Figs. 3 e 4 mostram o resultado da aplicação do *EM* para o conjunto de dados preparado com escore mínimo para alinhamento de 30%, e, nesse caso, percebe-se que o SID, pelo modelo de *Larsen* (Fig. 3), organizou e separou, como previsto, os elementos em três grupos, isto é, os grupos *SNP confirmado*, cujos elementos se encontram concentrados na parte superior do gráfico; *SNP não confirmado*, com os elementos concentrados na parte intermediária; e, na parte inferior do gráfico, estão os elementos do grupo *SNP descartado*. Entretanto, pelo modelo de *Mamdani* (Fig. 4), não foi obtido tanto sucesso, pois não existe distância suficiente entre os grupos, que caracterize uma separação nítida entre os mesmos. Para o modelo de *Larsen*, o grupo *SNP confirmado* foi composto de 5% dos elementos avaliados e, em relação ao modelo de *Mamdani*, esse mesmo grupo foi composto de 17% desse mesmo total.

Esse caso apresentado nas Figs. 3 e 4 trata-se de um caso de sucesso da aplicação do modelo de mineração e deve ser interpretado considerando os seguintes aspectos:

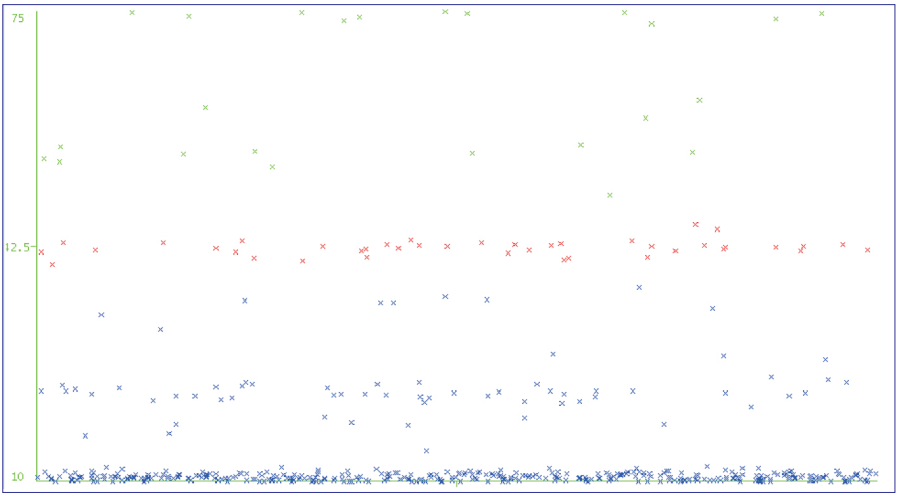

Fig. 3. Resultado do agrupamento dos dados determinados pelo SID, segundo o modelo de Larsen, para o conjunto de alinhamentos gerados com escore mínimo de 30%.

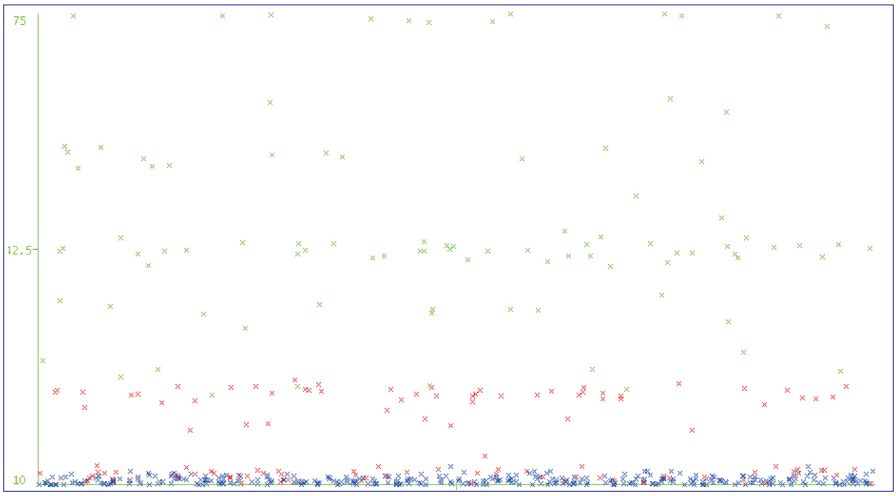

Fig. 4. Resultado do agrupamento dos dados determinados pelo SID, segundo o modelo de Mamdani, para o conjunto de alinhamentos gerados com escore mínimo de 30%.

a) Os eixos das abscissas e ordenadas nas Figs. 3 e 4, assim como, nas subsequentes (Figs. 5, 6, 7, 8, 9, 10, 11, 12, 13 e 14), representam, respectivamente, as posições possivelmente polimórficas, ou seja, o total de candidatos a SNP, e o valor inferido que permite particionar os candidatos. Isto é, o valor do eixo das ordenadas permite estratificar o candidato em relação aos três agrupamentos propostos no modelo: *SNP confirmado*, *SNP descartado* e *SNP não confirmado*;
b) A amplitude e os valores dos eixos das ordenadas são valores de referência atribuídos em consequência da função de saída e do método de *defuzzyficação* definidos no SID, portanto, não possuem interpretação quantitativa. A interpretação que deve ser feita desses valores é que, como consequência desses mesmos elementos do SID, quanto maior for o valor estabelecido para o candidato a SNP, maior será a sua caracterização como elemento do grupo *SNP confirmado* e, no sentido oposto, quanto menor for o valor estabelecido para o candidato a SNP, maior será a sua caracterização como elemento do grupo *SNP descartado*;
c) A interpretação desses resultados é, portanto, qualitativa e deve ser considerada a distribuição dos elementos em seções longitudinais – paralelas ao eixo das abscissas – onde, em um caso ideal, deve ser obtido, como anteriormente informado, a "menor distância" entre os elementos de um grupo e a "maior distância" entre os grupos. A Fig. 3 é um bom exemplo para o caso ideal. Contudo, a Fig. 4, apesar de apresentar os elementos particionados em três grupos, que é o objetivo do modelo proposto e podem ser diferenciados pelas cores, não apresenta seções longitudinais que facilmente possam delimitar e separar os grupos entre si;
d) Um exemplo que não se aproxima do caso ideal, em relação ao modelo proposto, pode ser visto na Fig. 5. Como será explicado a seguir, para o conjunto de dados representado nessa figura, o algoritmo *EM* deveria obter seis grupos, entretanto, particionou os elementos em sete grupos, como pode ser identificado pelas cores dos elementos, e, ainda, não foram obtidas a menor distância entre os elementos e nem a maior distância entre os grupos. Isto é, os elementos apresentam-se misturados e dispersos ao longo de todo o gráfico.

Para que fosse possível estabelecer um contraste entre os valores de probabilidade determinados pelo *Polyphred* e pelo *Polybayes* em relação ao resultado obtido no SID, utilizando o mesmo conjunto de dados e o mesmo algoritmo, foram agrupados os atributos *PPS* e $PPS_{Polybayes}$, para verificar o comportamento dessas medidas, uma vez que estabelecem as seis classes definidas pelo *Polyphred score*. Entretanto, como pode ser visto nas Figs. 5 e 6, esses dados não se agruparam de forma semelhante às classes que determinam.

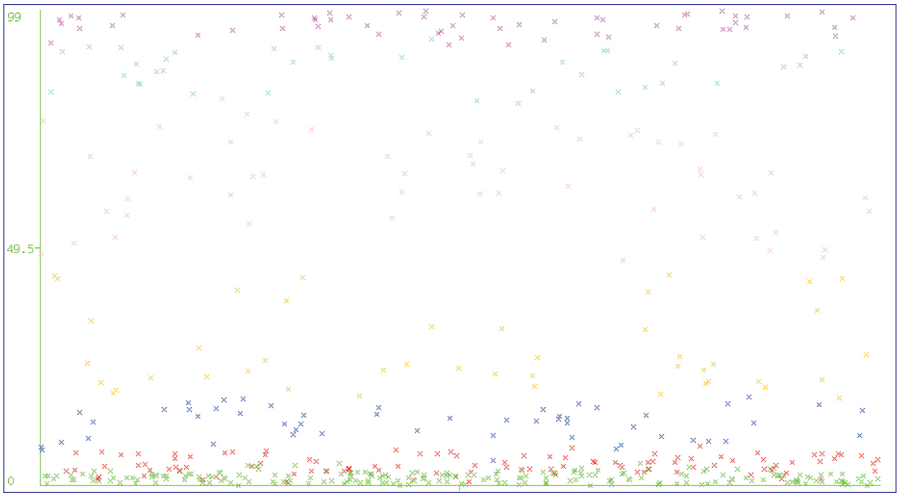

Fig. 5. Resultado do agrupamento dos dados de probabilidade de SNPs determinados pelo Polyphred, para o conjunto de alinhamentos gerados com escore mínimo de 30%.

Para o conjunto de sequências preparadas com escore mínimo de alinhamento em 60%, o SID não foi capaz de determinar três grupos distintos, de acordo com o algoritmo *EM*, como pode ser visto nas Figs. 7 e 8. A aplicação do SID com o modelo de Larsen foi relativamente melhor, se comparada ao resultado obtido com o modelo de *Mamdani*, pois, apesar de determinar apenas dois grupos, a separação deles foi bem caracterizada. Contudo, o resultado obtido com a aplicação do algoritmo *EM* sobre a inferência pelo modelo de *Mamdani* identificou cinco grupos, sendo que, entre dois desses grupos, não foi possível

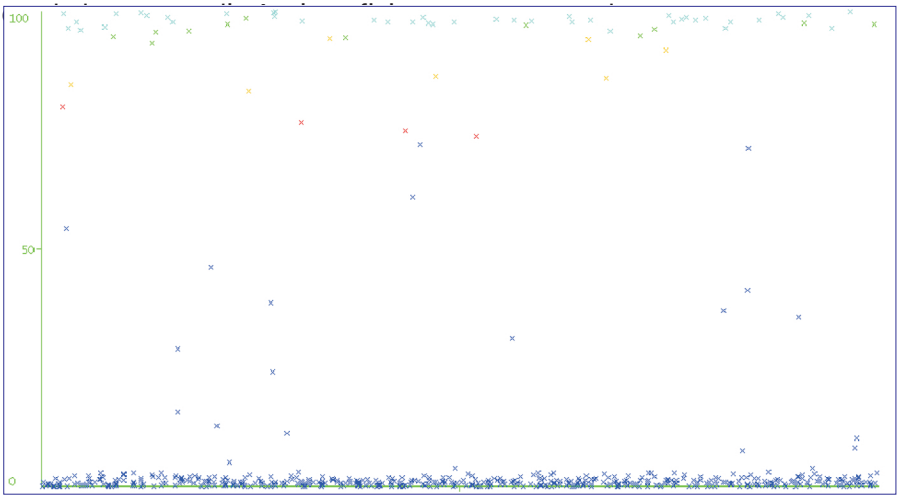

Fig. 6. Resultado do agrupamento dos dados de probabilidade de SNPs determinados pelo Polybayes, para o conjunto de alinhamentos gerados com escore mínimo de 30%.

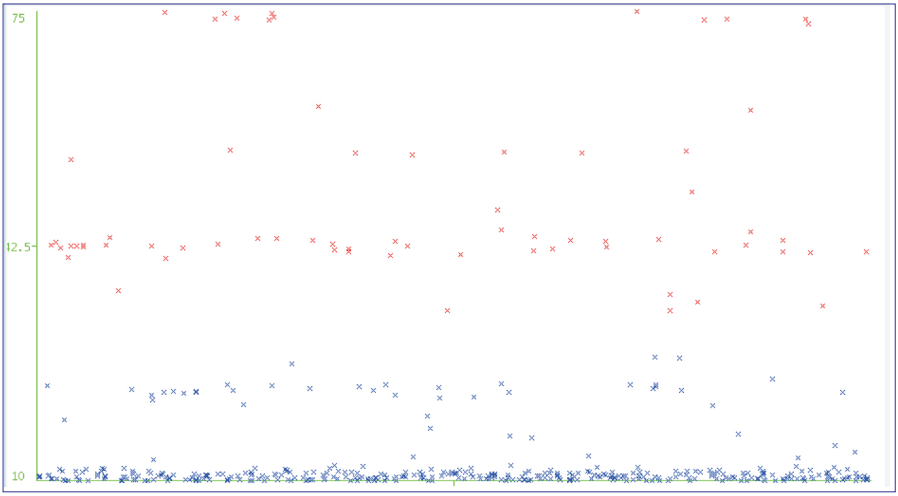

Fig. 7. Resultado do agrupamento dos dados determinados pelo SID, segundo o modelo de Larsen, para o conjunto de alinhamentos gerados com escore mínimo de 60%.

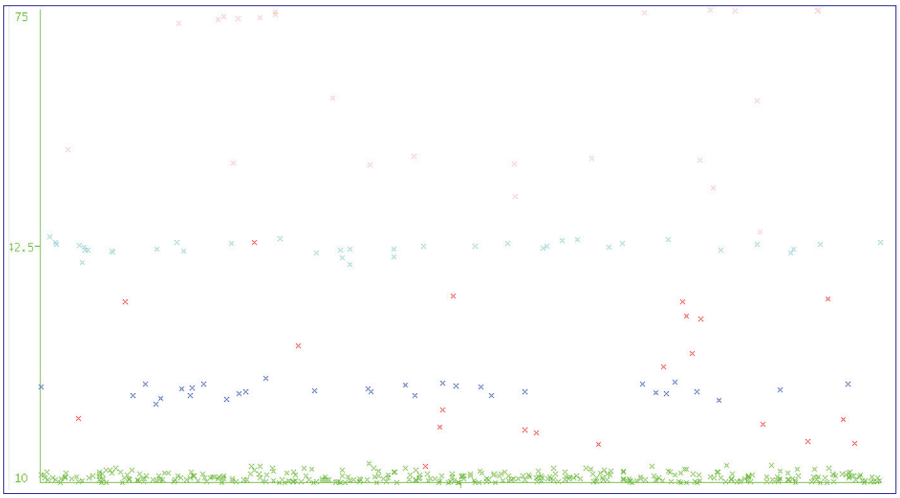

Fig. 8. Resultado do agrupamento dos dados determinados pelo SID, segundo o modelo de Mamdani, para o conjunto de alinhamentos gerados com escore mínimo de 60%.

Além disso, o agrupamento dos valores de probabilidade determinados pelo *Polyphred* e pelo *Polybayes*, com esse mesmo algoritmo e sobre esse mesmo conjunto de dados, também não apresenta o resultado esperado, sendo que os agrupamentos não refletem as classes que representam, como pode ser visto nas Figs. 9 e 10.

A verificação do terceiro conjunto de sequências, que foram obtidas com escore mínimo para alinhamento de 90%, apresentou o melhor resultado sob os dois modelos de inferência, discriminando os três grupos esperados e, segundo o modelo de *Larsen* (Fig. 11), estabelecendo o grupo *SNP confirmado* com 9% dos elementos avaliados e, em relação ao modelo de *Mamdani* (Fig. 12), esse mesmo grupo apresentou de 10% desse mesmo total.

Contudo, a exemplo do que ocorreu com os conjuntos de dados anteriores, o agrupamento dos valores de probabilidade determinados pelo *Polyphred* e pelo *Polybayes* também não apresentou o resultado espera-

do. Como pode ser visto nas Figs. 13 e 14, esses dados não se agruparam de forma semelhante às classes que determinam.

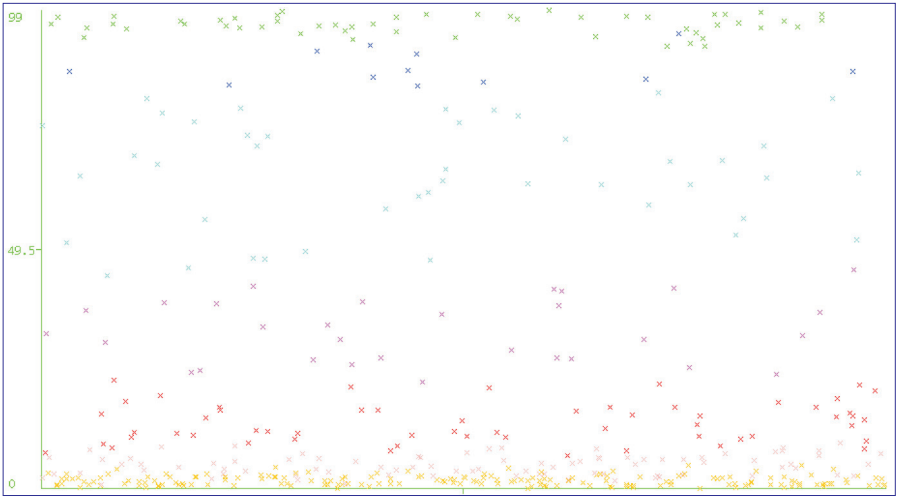

Fig. 9. Resultado do agrupamento dos dados de probabilidade de SNPs determinados pelo Polyphred, para o conjunto de alinhamentos gerados com escore mínimo de 60%.

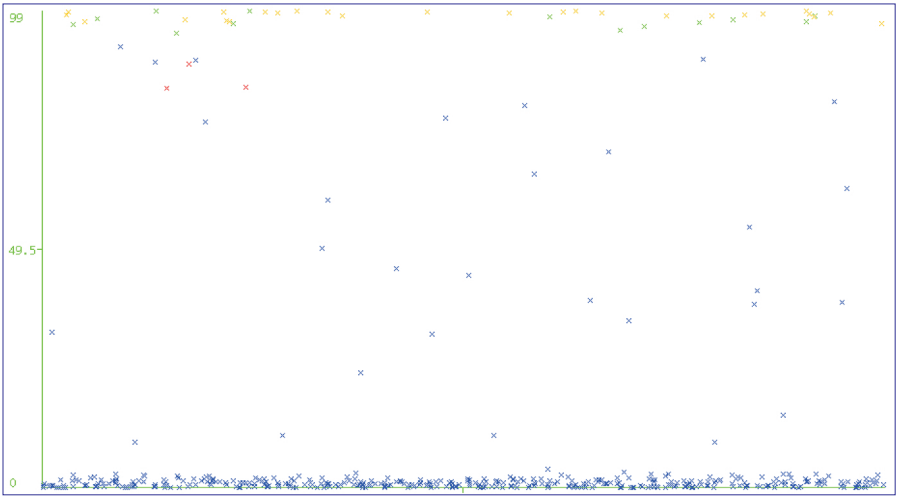

Fig. 10. Resultado do agrupamento dos dados de probabilidade de SNPs determinados pelo Polybayes, para o conjunto de alinhamentos gerados com escore mínimo de 60%.

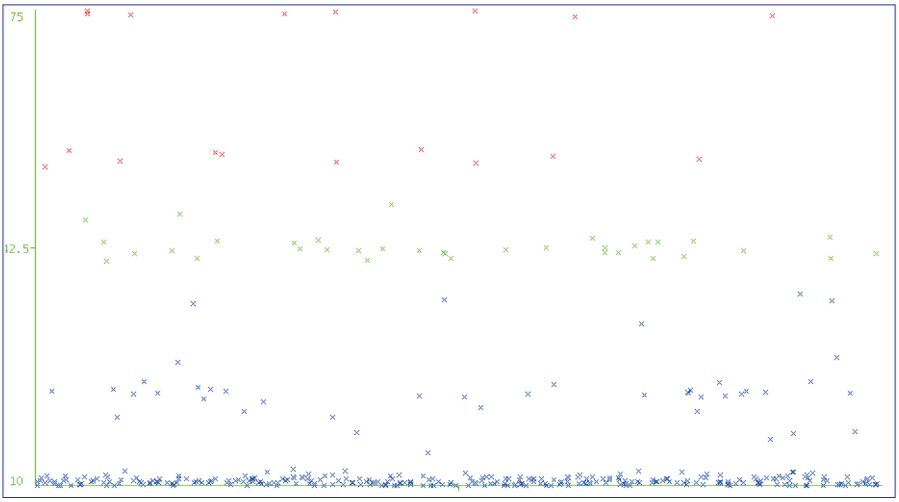

Fig. 11. Resultado do agrupamento dos dados determinados pelo SID, segundo o modelo de Larsen, para o conjunto de alinhamentos gerados com escore mínimo de 90%.

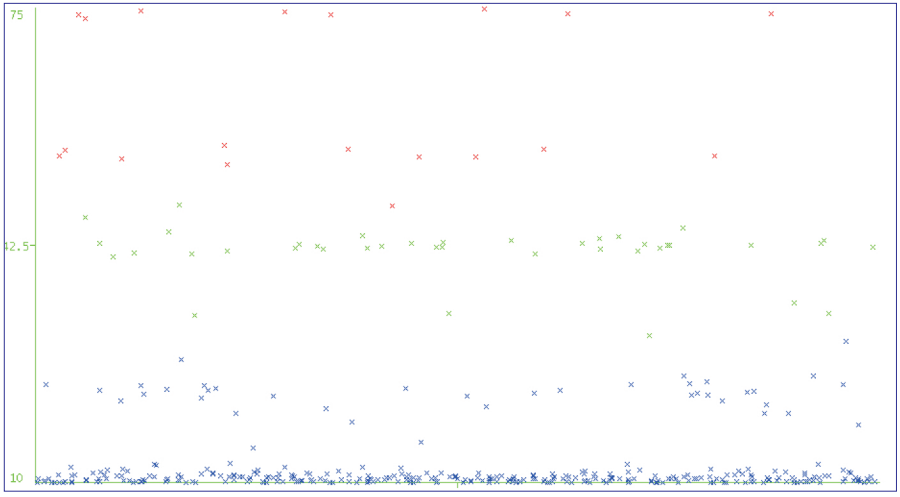

Fig. 12. Resultado do agrupamento dos dados determinados pelo SID, segundo o modelo de Mamdani, para o conjunto de alinhamentos gerados com escore mínimo de 90%.

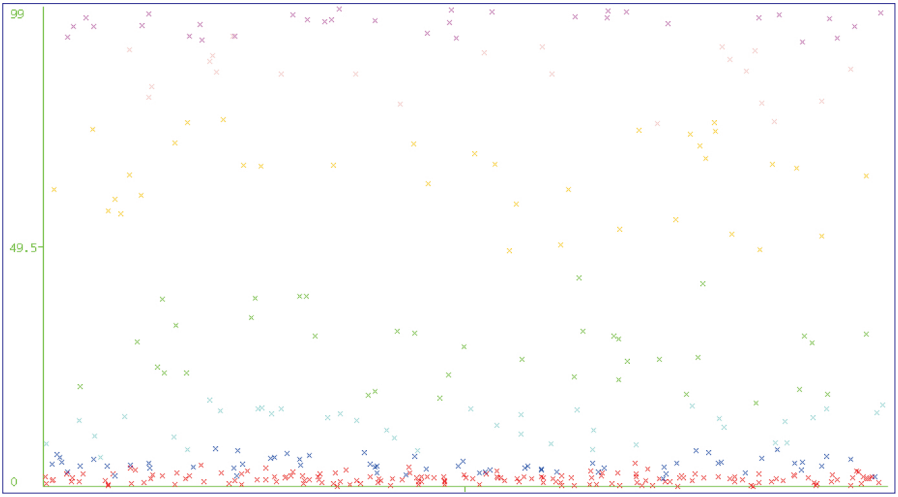

Fig. 13. Resultado do agrupamento dos dados de probabilidade de SNPs determinados pelo Polyphred, para o conjunto de alinhamentos gerados com escore mínimo de 30%.

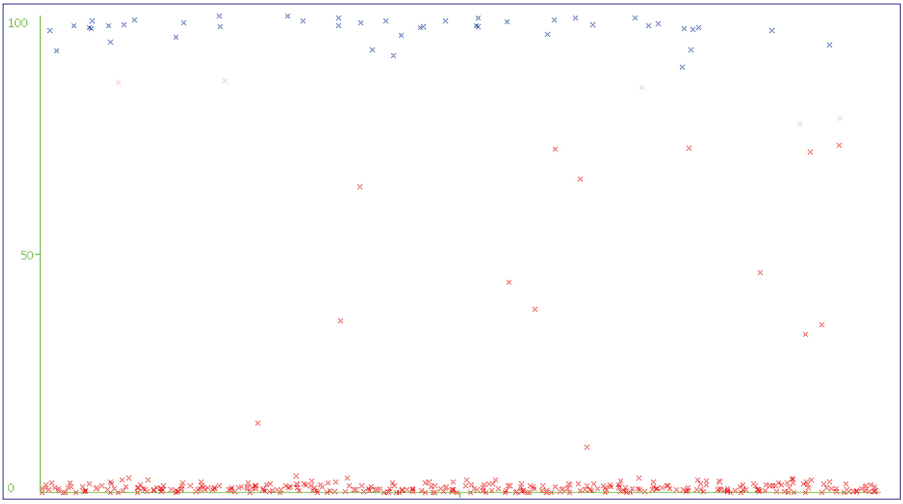

Fig. 14. Resultado do agrupamento dos dados de probabilidade de SNPs determinados pelo Polybayes, para o conjunto de alinhamentos gerados com escore mínimo de 90%.

Alguns dos resultados obtidos com essa análise podem ser vistos na Tabela 6, que traz o percentual, em relação ao total de candidatos a SNPs, quantificado na Tabela 3, estratificados de acordo com a classificação proposta pelo modelo.

Tabela 6. Distribuição dos candidatos a SNP determinada pelo SID, pelos modelos de inferência de Mamdani e Larsen, conforme o modelo proposto.

Escore mínimo de alinhamento	Candidatos a SNP	SID - Mamdani			SID - Larsen		
		SNP_C	SNP_{NC}	SNP_D	SNP_C	SNP_{NC}	SNP_D
30%	1.624	17%	22%	61%	9%	4%	87%
60%	1.378	-	-	-	-	-	-
90%	1.145	10%	5%	85%	9%	5%	86%

Conclusões

Um dos resultados dessa pesquisa foi o desenvolvimento, a implantação e aplicação de um modelo computacional para identificação de informação genômica associada à resistência ao carrapato bovino sobre um conjunto de dados que, até então, não havia sido analisado e, até a presente data, foi analisado sob outros aspectos, mas não por modelos computacionais. Assim, esse projeto constitui a primeira análise dessas informações por modelos computacionais, especificamente um modelo de mineração de dados, sob técnicas de aprendizado de máquina, implementado por um SID e a partir de uma metodologia de investigação própria que também foi desenvolvida dentro do escopo desse trabalho.

A proposta de desenvolver um SID para esse fim, entre outras razões, também foi determinada pelas características dos resultados obtidos com ferramentas usuais de identificação desse tipo de polimorfismo, nas situações em que:

a) critérios fixos e precisos de classificação não são adequados, quando a análise dos candidatos a SNPs apresenta resultados em que os candidatos a SNPs situam-se próximos à divisão das classes;
b) as ferramentas usuais de identificação apresentam resultados conflitantes.

Assim, tomando como base os fundamentos dos conjuntos difusos e da lógica difusa, foi possível a implementação de um modelo de inferência que permita a análise e o tratamento de resultados prévios ou intermediários, frente a características de imprecisão e incerteza, que são comuns em processos decisórios.

O modelo de inferência foi desenvolvido com a utilização do *fuzzyMorphic.pl*, uma ferramenta de implementação de SIDs, desenvolvida como parte dessa pesquisa, e que permite a implementação de SIDs para diversos fins, bastando, para isso, descrever os valores *crisps* de entrada, as funções de pertinência, as regras de inferência, máquina de inferência e a função de saída.

O SID foi aplicado sobre o conjunto de dados para identificar SNPs como parte do procedimento de mineração de dados, isto é, antes da investigação propriamente dita, a metodologia estabelece a execução do protocolo planejado, que inclui a preparação dos dados de entrada, composto das seguintes etapas:

1) geração das sequências, a partir dos cromatogramas e das sequências-consenso;
2) obtenção de resultados prévios com a identificação de candidatos a SNPs por ferramentas usuais;
3) seleção, extração, composição e organização dos resultados prévios em um conjunto de dados de entrada para o SID;
4) mineração dos dados por meio do SID;
5) análise dos resultados obtidos com a aplicação do SID.

Além do modelo matemático e computacional para a investigação de SNPs, entre as contribuições desse trabalho, destacam-se o desenvolvimento do *fuzzyMorphic.pl* e a sua forma de utilização, à medida que:

a) quanto ao desenvolvimento do *fuzzyMorphic.pl*, ressalta-se que como não foi implementado especificamente para o problema em questão, pode ser utilizado no desenvolvimento de SIDs para variados problemas e modelos de inferência difusa;

b) com relação à sua forma de utilização, a facilidade com que permite a descrição de um modelo, possibilita que um estudo seja realizado sob diferentes condições, o que é particularmente necessário e importante para os procedimentos de pesquisa.

Referências

ARBEX, W. A. **Modelos computacionais para identificação de informação genômica associada à resistência ao carrapato bovino.** 2009. 200 f. Tese (Doutorado em Engenharia de Sistemas e Computação) – Universidade Federal do Rio Janeiro, Rio de Janeiro, RJ.

ATHAYDE, A. C. R.; FERREIRA, U. L.; LIMA, E. Á. L. A. Fungos entomopatogênicos: uma alternativa para o controle do carrapato bovino *Boophilus microplus.* **Biotecnologia, Ciência & Desenvolvimento**, v. 4, n. 21, p. 12-15, jul. 2001.

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

MARTINEZ, M. L.; SILVA, M. V. G. B. da; MACHADO, M. A. et al. A biologia molecular como aliada no combate aos carrapatos. In: SIMPÓSIO DA SOCIEDADE BRASILEIRA DE MELHORAMENTO ANIMAL, 5., 2004, Pirassununga. **Anais...** Pirassununga: SBMA, 2004. Disponível em: <http://sbmaonline.org.br/anais/v/palestras/pdfs/palest13.pdf>. Acesso em: 15 jul. 2009.

NICKERSON, D. A.; TAYLOR, S. L.; KOLKER, N. et al. **Polyphred users manual**. Seattle: University of Washington, 2008. Version 6.15 Beta.

TEODORO, R. L.; MARTINEZ, M. L.; SILVA, M. V. G. B. da et al. Resistência bovina ao carrapato *Boophilus microplus*: experiência brasileira. In: Anais do V SIMPÓSIO DA SOCIEDADE BRASILEIRA DE MELHORAMENTO ANIMAL, 5., 2004, Pirassununga. **Anais...** Pirassununga:

SBMA, 2004.Disponível em: <http://sbmaonline.org.br/anais/v/palestras/pdfs/palest12.pdf>. Acesso em: 15 jul. 2009.

VAZ, I. S.; TERMIGNONI, C.; MASUDA, A. et al. Vacina contra carrapato: *Boophilus microplus*: controlando o carrapato com o uso de vacina. **Biotecnologia, Ciência & Desenvolvimento,** v. 2, n. 13, p. 18-23, mar. 2000.

WITTEN, I. H.; FRANK, E. **Data mining:** pratical machine learning tools and techniques. 2. ed. San Francisco: Morgan Kaufmann Publishers, 2005.