Topic: **Text Mining Information Extraction**

SNP DISCOVERY IN CDNA SEQUENCES WITH FUZZYMORPHIC.PL

W Arbex[1], LA Carvalho[2], MV Silva[1]

[1]*Empresa Brasileira de Pesquisa Agropecuária*
[2]*Universidade Federal do Rio de Janeiro*

Differences between specific base pairs of different aligned sequences are the most common type of generic variability. Such differences, know as single nucleotide polymorphisms (SNPs), are important in the study of variability of species, because they may cause functional or phenotypic changes, which, by their turn, can result in evolutionary or biochemical effects on the individuals of the species. The use of computational algorithms to SNPs investigation is a widespread practice and the Polyphred and Polybayes programs stand out, because they are widely used. Thus, it is expected these programs show similar results when they are using the same data set, despite used different methods, but it isn't unusual show conflicting results. The PhD thesis *"Computational models to the identification of genomic information associate to the resistance to cattle tick"*, propose a fuzzy inference model to aid in the process decision, using the fuzzyMorphic.pl, a computational tool, write in Perl language, which allow the modeling and implementation of fuzzy inference system to data mining and this text present the inference fuzzy model, proposed and developed in context of quoted thesis, which was implemented with the fuzzyMorphic.pl, to aid decision support from previous or primary results to SNPs discovery in cDNA sequences by data mining, using machine learning techniques, fuzzy logic specifically. The model proposed implements a data mining methodology to explore these results to aid decision support when the results are divergent or confirm them when they are similar. The methodology is based in a two steps protocol to data mining: the data pre-processor section and the fuzzy inference section. The data pre-processor section provides extraction, integration, selection, completion and deletion procedures, which are the common procedures to pre-processing for data mining and, depending on the characteristics of research, requirements for the processing of data. The second section brings the fuzzy system itself, which includes the fuzzification, inference and defuzzification procedures. The fuzzification processes can be described by membership functions composed by fuzzy sets in a standard format, to the inference process can be used the Mamdani's or Larsen's inference models and the defuzzification process can be represented by an output function with fuzzy sets in standard format and, furthermore, using the "center of maximum" as the defuzzification method, because it takes "multiply shots" on the output function. The Polyphred's method search for positions in sequences - "reads" - where were detected more than one nucleotide and the Polybayes' method look for polymorphic sites by evaluating the different nucleotides within cross-sections of a multiple alignment. However, both methods do not consider the base quality in the sequence consensus resulting of the alignment. The described model in this text combines this base quality with the previous results obtained from the Polybayes and Polyphred, setting new attributes to SNPs identification.

5TH INTERNATIONAL CONFERENCE OF THE BRAZILIAN ASSOCIATION FOR BIOINFORMATICS AND COMPUTATIONAL BIOLOGY

A B
B
B C

X-meeting

October 18 to 22, 2009

Hotel do Frade - Angra dos Reis - Rio de Janeiro - Brazil

X-Meeting
Abstracts
Book
2009

Angra dos Reis
Rio de Janeiro
Brazil