Methodological Procedure for Decision-Making Using Fuzzy Inference for SNP Discovery

Wagner Arbex[1], Marta Martins[1], Marcos Vinícius Silva[1] and Luis Alfredo Carvalho[2]
[1]Brazilian Agricultural Research Corporation, Juiz de Fora, MG, Brazil
[2]Federal University of Rio de Janeiro, Rio de Janeiro, RJ, Brazil

Abstract— *Problems when dealing with imprecise or uncertain features, e. g., problems of decision-making, can be designed as fuzzy systems, since these systems allow subjective and qualitative arguments, which are usually intrinsic in such problems, to be processed. Research involving the discovery of single nucleotide polymorphisms (SNPs) requires bioinformatics tools to be applied to different cases with an ability to analyze "reads" from different sources and levels of coverage and also to establish reliable measures. These tools work with different methodologies in regards to distinct attributes. When dealing with the same data set, similar results are expected. However, sometimes such different methodologies may yield different results, which leads to uncertainty in the decision-making process. This paper presents a methodology based on the fuzzy inference decision model applied to bioinformatics, based on results from two other tools for SNPs discovery.*

Keywords: Fuzzy inference, decision support, single nucleotide polymorphism, SNP, SNP discovery

1. Introduction

Data generation technologies for molecular biology challenge the development of appropriate computer systems and require accurate bioinformatics tools for analyzing such data. In this sense, machine learning appears as a promising alternative for knowledge discovery in genomic databases, using both decision-making and data mining techniques, among other resources of artificial intelligence.

In the fields of genomics and bioinformatics, the already great amount of data continues to grow very quickly, widening the gap between the generation and interpretation of such data. Therefore, different ways to reduce the problem of huge quantities of data as opposed to the ability to interpret them are studied. For instance, fuzzy inference systems implement computational models for data mining aimed at discovering knowledge in databases. Such models are capable of processing imprecise and qualitative information and, therefore, they are suitable in situations that require decision-making [1].

This paper aims at describing a computational model that uses fuzzy logic as the basis for the implementation of an inference system aimed at assisting decision-making. More information about the inference model proposed and its applications can be found in the research project "Computational models for the identification of genomic information associated to the resistance to cattle tick" [2].

In support to such description, the concept of single nucleotide polymorphism (SNP) and the use of fuzzy inference to deal with uncertainty, imprecision and decision-making problems will be presented. Following, the fuzzy inference model and the methodological approach will be presented and discussed. They work on previous results obtained by different SNP discovery tools that have possibly conflicting results; therefore, the methodology is applied to assist decision-making in cases when information is conflicting and also in the confirmation of coincident information.

2. Background

2.1 Single Nucleotide Polymorphisms

Sequencing projects have shown that genomes have more variations and more complexity than initially expected. One of such variations and peculiarities are the SNPs, that is, base pairs in a single position in genomics DNA that are presented in sequences with different alternatives [3]. SNPs can be found in the genome of a single individual or groups of individuals, in a given population (Fig. 1).

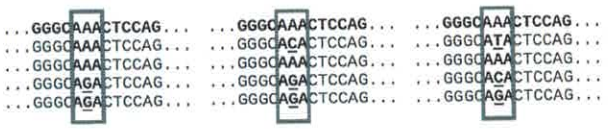

Fig. 1: Hypotetical instances of SNPs bi, tri and tetra-allelic, respectively. The first line, in bold, shows the consensus sequence and the underlined bases are the SNPs. Actually, the ocurrence of bi-allelics SNPs it's not only more common, but almost absolute in relation to the others [4].

Individuality is a result of genetic expression, that is, in essence, the nucleotide sequences form DNA and RNA, as well as protein sequences, which interact and, in turn, form cells, which also interact and form tissues, organs, until, eventually, make individuals. In this relies the importance of SNPs: if a single nucleotide, a single base in a given sequence, is changed, it may alter the formation of proteins and, altogether, these changes may cause variations in the individuals.

2.2 Fuzzy Inference Approach

Classical approaches are insufficient to analyze values very close to the limits of a given category; therefore, one may get results that are questionable, though mathematically and logically accurate. For instance, the Polyphred Score (PPS) [5] determines six classes with precise intervals (Tab. 1). Assuming that the scores 70 and 89 were taken for two points, respectively, then, when deciding whether these two points are SNPs, a 35% of true positives rate (Rank 4) would be considered for both.

Table 1: Accuracy by PPS and rank.

Rank	PPS	True positives rate
1	99	97%
2	95 – 98	75%
3	90 – 94	62%
4	70 – 89	35%
5	50 – 69	11%
6	0 – 49	1%

This logically and mathematically precise decision can be questioned because of the subjectivity involved. Both scores, 70 and 89, are very close to the limits of the classes to which they belong, and, therefore, different interpretations are supported for these scores. However, traditional approaches to logics and mathematics do not have the necessary tools to handle threshold values, or even imprecision or uncertainty. Specifically, threshold values result in doubt when it comes do deciding whether a given base is polymorphic or non-polymorphic, which suggests a fuzzy inference system for handling this uncertainty.

Usually, the problem with threshold values is not as simple as it may seem, if it were so, classical approaches could easily solve it. However, the closer to the subjective reasoning for the interpretation and the extraction of an answer or a decision, the more complex it becomes and the apparent simplicity is given by fuzzy logic modeling and by its basis in the theory of fuzzy sets.

3. Decision-Making with Fuzzy Inference

The subjectivity inherent to reasoning is capable of dealing with complex situations, based on inaccurate, uncertain or approximate information and, therefore, the strategy is to use human operators of an also imprecise nature, which are expressed in linguistic terms or variables. In order to describe or handle problems, such essentially human proposal, generally, does not generate a solution in terms or exact numbers, but, for instance, leads the solution to a qualitative classification, clustering or aggregating results into categories or possible solutions sets [1]. These solutions can be seen as a result of the "principle of incompatibility" [6].

The linguistic terms or variables increase the complexity of traditional models and computational systems concerning their ability to handle exact numbers and discrete values – which are, sometimes, mutually exclusive. Hence, working with uncertain values may enable the modeling of complex systems, even if they reduce the accuracy of the result, without, thought, leading to loss of credibility.

If uncertainties, when viewed in isolation, are undesirable, when they are associated with other characteristics, they generally allow the reduction of system complexity and increase the credibility of the results [7].

Fuzzy sets theory and fuzzy logics are appropriate to represent, in mathematical terms, the inaccurate information that can be expressed by a set of linguistic rules. Also, if there is the possibility for human operators to be organized as a set of conditional statements (in the if ANTECEDENT then CONSEQUENT form); thus, subjective reasoning can be expressed in the form of computationally executable algorithms [8] [6] with the ability for imprecisely classifying the antecedent and consequent variables of conditional statements as qualitative (instead of quantitative) concepts, which represents the idea of a linguistic variable [1].

Hence, since they are capable of efficiently processing inaccurate and qualitative information, fuzzy inference models are suitable in situations that require decision-making [1].

4. Fuzzy Inference model for identification of SNP candidates

4.1 Methodological Procedure with Fuzzy Inference System for Decision-Making

The function structure of the machine learning model for decision-making is represented inFig. 2 and 3, in which there is emphasis on the division of the system's workflow in well-defined stages:

1) initial processing of the chromatograms, when bases are read, and, consequently, sequences ("reads") and contiguous sequences are originated and, besides, the quality of the bases of these sequences is determined. This stage is done by phredPhrap pipeline [9], and many files are generated, such as the "ace" file and several "phd" files, from each sequence read (Fig. 2);
2) Polyphred [10] and Polybayes [11] software run on "ace" and "phd" files, and each of these programs, following its own methodology, identifies the SNP candidate bases and determines a probability for each of these bases. These results are recorded in "polyphred.out" and "report.out" files, which will be used as input to the learning procedure (Fig. 2);
3) in this next stage, preparation of the data is carried out. Data from Phrap [9] – generated by the phredPhrap pipeline – Polyphred and Polybayes are extracted and selected from their respective files and, if necessary, they are complemented. This stage of preparing the data is done by parsepolyBayes.pl, parsepolyPhred.pl, parsephrapQality.pl and joinparsersOut.pl scripts [2].

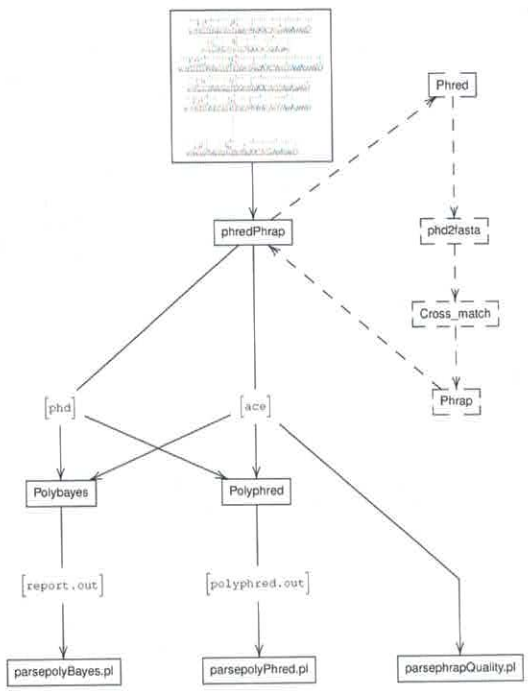

Fig. 2: Synthesis of the functional structure of the model of machine learning (I).

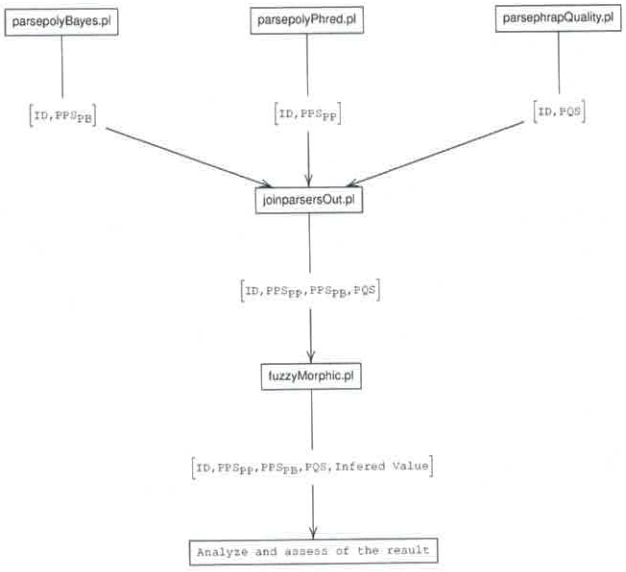

Fig. 3: Synthesis of the functional structure of the model of machine learning (II).

Furthermore, the joinparsersOut.pl script forms the file in a specific structure to fuzzyMorphic.pl [12] software (Fig. 3);

4) while running fuzzyMorphic.pl, the machine learning procedure is performed, implementing a fuzzy inference system to make an output file with the same input data, adding the inferred value about the investigated feature (Fig. 3);
5) in order to analyze and assess the outcome, we use certified techniques and tools so as to check the inferred results. In this case, a cluster analysis is carried out in the resulting data set, which arises from the fuzzy inference system (Fig. 3).

4.2 Review and Discussion of the Methodological Procedure

The machine learning model implemented, functionally speaking, explores the data set which was created by connecting Polyphred and Polybayes output data sets. Then, it checks the probabilities for each element of this data set, as specified by their different proposals. Next, this model defines, for each element, a new attribute, which should be used as a reference in the attempt to cluster data set into groups of elements that can be seen as confirmed polymorphic points (SNP confirmed), non-polymorphic points (SNP discarded), and also points without sufficient evidence for a conclusive definition (SNP not confirmed).

However, any classification one could propose might be influenced by the data "form" or "behavior". Also, in regards to classes defined by exact limits, questionable decisions may arise when the value is very close to the limits of classes. These issues, among others, suggest the adoption of non-hierarchical and non-supervised partitioning methods, because these methods do not refer to any external premises to establish the classes that may divide a set, but, rather, its premises are established by specific features, which are internal and inherent to the data set evaluated. Therefore, the adoption of these methods removes or reduces the action of external agents, such as a priori definition of precise limits for the classes, on the model.

Premises of partitioning methods from non-hierarchical algorithms are based on their own set of values assessed, searching for maximum internal cohesion of a group of objects and for maximum detachment between groups [13]. From another perspective, analyzing the set itself, they try to identify the elements that, concerning the attribute evaluated, are closer to the other elements of the group, and, once the groups are established, the elements with a given feature should be as far as possible from the elements belonging to the elements in another group. Thus, as these premises are due to their own values analyzed, the effect of data behavior is reduced, that is, assuming that the attribute evaluated presents a certain trend, all elements have the same behavior and an undirected partitioning taken from elements

themselves can reduce or eliminate this tendency.

The exclusion of external premises as well as the reduction of the models adopted for assessing the results can be advantageous, insofar as they simplify the answers, reducing the risk of them being manipulated. If possible, these models should be self-contained, independent of external components and use as few variables and parameters as possible, avoiding "boundary conditions", which enable the "accommodation" of a result, instead of truly finding it.

Determination of data clustering is a complex and hard to implement task, because it is necessary to find out how the data are and into how many classes the data are distributed, without any previous knowledge about them. Classes may not even exist, if the elements are distributed equitably over the space and do not feature any category, for clusters or classes are based on the similarity between elements. Eventually, the verification of resulting classes is performed so as to assess whether there is some sort of useful meaning [13].

Following this analysis, the model implemented from machine learning techniques replaces, through fuzzy inference, a continuous probability measure in the interval [0,1] associated with the probability of the point becoming an SNP, by another attribute, which allows clustering the points into three partitions: SNP confirmed, SNP discarded and SNP not confirmed. Thus, after data processing by the fuzzy inference system, which aims at clustering the resulting data through a non-supervised algorithm and dynamically establishing the number of groups, hoping that the result obtained confirms the partitioning of the set into three groups based on the new attribute.

Operationally, this procedure is done by fuzzyMorphic.pl software, which implements the fuzzy inference system and determines this new attribute, while the clustering analyses are aided by Weka (Waikato Environment for Knowledge Analysis) [14] software.

Among clustering algorithms, Weka implements the Expectation-Maximization (EM) algorithm, which has the feature of determining, in runtime, the number of clusters which fits better the elements analyzed, without this information being previously provided to it. EM algorithm was developed for statistical inference problems in general, and it seeks to locate the value for a parameter that maximizes the likelihood function. For the clustering procedure, the data standard division was adopted, that is, 2/3 and 1/3 for training and testing, respectively.

5. Conclusion

Generally, fixed and precise criteria of classification are not suitable when studies show results very close to a certain limit, for instance, a classes division. Nevertheless, these cases can be approached by fuzzy inference systems, which are also convenient, as well as able to handle uncertain and imprecise problems in decision-making.

When adding a new attribute to previous results, the fuzzy system is able to decide, uniquely among the three possibilities resulting from the model, and then it clusters them through a non-supervised algorithm with dynamic establishment of the number of clusters, hoping that the outcome of this clustering confirms set partitioning into three clusters, and requiring no fixed and/or precise limits to classify and, thus, identify potential SNPs.

Acknowledgements

The authors would like to express thanks to the State of Minas Gerais Research Support Agency (Fapemig) for the partial support for the accomplishment of this paper.

References

[1] P. E. M. de Almeida and A. G. Evsukoff, "Sistemas fuzzy," in *Sistemas inteligentes: fundamentos e aplicações*, S. O. Rezende, Ed. Barueri: Manole, 2005, pp. 169–202.

[2] W. Arbex, "Computational models for the identification of genomic information associated to the resistance to cattle tick," Systems Engineering and Computer Science Program, PhD thesis, Federal University of Rio de Janeiro, Rio de Janeiro, 2009.

[3] A. J. Brookes, "The essence of SNPs," *Gene*, vol. 2, no. 234, pp. 177–186, 1999. DOI: 10.1016/S0378-1119(99)00219-X.

[4] T. Brown, *Genomes*, 2nd ed. New York: John Wiley & Sons, 2002.

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

[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. DOI: 10.1109/TSMC.1973.5408575. [Online]. Available: http://www-bisc.cs.berkeley.edu/Zadeh-1973.pdf.

[7] G. J. Klir and B. Yuan, *Fuzzy sets and fuzzy logic: theory and applications*. Upper Saddle River: Prentice Hall, 1995, p. 592, ISBN: 0131011715.

[8] R. Tanscheit, "Sistemas fuzzy," in *Inteligência computacional: aplicada à administração, economia e engenharia em Matlab*, H. A. e Oliveira Júnior, Ed. São Paulo: Thomson Learning, 2007, pp. 229–264.

[9] P. Green, *Phrap*, 1 CD, C. Linux environment with C compiler., 1999. [Online]. Available: http://www.phrap.org/index.html.

[10] 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. DOI: 10.1093/nar/25.14.2745. eprint: http://nar.oxfordjournals.org/cgi/reprint/25/14/2745.pdf.

[11] 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, 1999. DOI: 10.1038/70570.

[12] W. Arbex, *fuzzyMorphic.pl*, 1 CD, Perl. UNIX-like environment with GUI and Perl 5.0 interpreter or newer., Juiz de Fora, 2009.

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

[14] I. H. Witten and E. Frank, *Data mining: pratical machine learning tools and techniques*, 2nd ed. San Francisco: Morgan Kaufmann Publishers, 2005, p. 525, ISBN: 0-12-088407-024884070.

PROCEEDINGS OF
THE 2012 INTERNATIONAL CONFERENCE ON
BIOINFORMATICS & COMPUTATIONAL BIOLOGY

BIOCOMP 2012

Editors

Hamid R. Arabnia
Quoc-Nam Tran

Associate Editors

Andy Marsh, Ashu M. G. Solo

WORLDCOMP'12
July 16-19, 2012
Las Vegas Nevada, USA
www.world-academy-of-science.org

This volume contains papers presented at The 2012 International Conference on Bioinformatics & Computational Biology (BIOCOMP'12). Their inclusion in this publication does not necessarily constitute endorsements by editors or by the publisher.

ISBN: 1-60132-204-6
Printed in the United States of America

CSREA Press
U. S. A.