Quasi Dominance Rough Set Approach in Testing for Traces of Natural Selection at Molecular Level

Krzysztof A. Cyran

Abstract. Testing for natural selection operating at molecular level has become one of the important issues in contemporary bioinformatics. In the paper the novel methodology called quasi dominance rough set approach (QDRSA) is proposed and applied for testing of balancing selection in four genes involved in human familial cancer. QDRSA can be considered as a hybrid of classical rough set approach (CRSA) and dominance rough set approach (DRSA). The advantages of QDRSA over CRSA and DRSA are illustrated for certain class of problems together with limitations of proposed methodology for other types of problems where CRSA or DRSA are better choice. The analysis of the reasons why QDRSA can produce decision algorithms yielding smaller error rates than DRSA is performed on the real world example, what shows that superiority of QDRSA in certain types of applications is of practical value.

Keywords: DRSA, CRSA, natural selection.

1 Introduction

Since the time of Kimura's famous book [7], the search for signatures of natural selection operating at the molecular level has become more and more important. It is so because neutral theory of evolution at molecular level does not deny the existence of selection observed at that level. It only states that the majority of observed genetic variation is caused by random fluctuation of allele frequencies in finite populations (effect of genetic drift) and by selectively neutral mutations.

If majority of mutations have been claimed to be neutral, then the next step should be to search for those which are not neutral. Indeed, several statistical tests, called

Krzysztof A. Cyran
Institute of Informatics, Silesian University of Technology,
Akademicka 16, 44-100 Gliwice, Poland
e-mail: krzysztof.cyran@polsl.pl

K.A. Cyran et al. (Eds.): Man-Machine Interactions, AISC 59, pp. 163–172.
springerlink.com

neutrality tests, have been developed and the neutral theory of evolution has been used as a null hypothesis for them. A statistically significant departure from this model can be therefore treated as a signature of natural selection operating in a gene under consideration.

Unfortunately, other reasons for departure from the neutral model are also possible and they also account for statistically significant signals in neutrality tests. These reasons include expansion of the population and geographical substructure of it with limited migration. Also recombination accounts for incorrect testing of the natural selection often suppressing the positive test signals even if the selection was present. Moreover, these effects affect various tests with diffrent strength, resulting in an interpretation puzzle instead of clear indication in the favor of the natural selection or against it.

Aforementioned difficulty in the interpretation of a battery of tests is the start point for machine learning methodology. The required assumption for successful application of machine learning is that a mosaic of test outcomes, making direct inference so troublesome, containes enough information to diferentiate between the existence of natural selection and its lack. The second prerequisite is the expert knowledge about presence of the selection for given combinations of neutrality test outcomes. Having those two it is possible in principle to train the knowledge retrieving system and, after successful testing, to use it for other genes for which the expert knowledge is unknown.

The paper presents a novel methodology called quasi dominant rough set approach (QDRSA) and compares its advantages and limitations with other rough set based methods: classical rough set approach (CRSA) and dominance based rough set approach (DRSA) using the aforementioned application. Such strategy, except for presentig theoretical aspects about the types of problems adequate for QDRSA, at the same time is able to demonstrate that the class of problems which can be solved with QDRSA is represented in real world applications similar to those considered in the paper.

2 Quasi Dominance Rough Set Approach

Since the first presentation of Rough Set Theory (RST) by Pawlak [8, 9] as an information retrieval system generating rules describing uncertain knowledge in a way alternative to Fuzzy Sets Methodology [12], many modifications of RST have been proposed. The most notable of them include Ziarko's Variable Precision Rough Set Model (VPRSM) [14], Dominance Rough Set Approach introduced by Greco, Matarazzo and Slowinski [5], and Near Set Approach (NST) developed by Peters [10].

The first is dedicated for large data sets where tolerated to some extent inconsistencies can be advanagous, the second is appropriate for attributes with inherent preference order and not necesserily discretized, and the latter uses the discovery of affinities between perceptual objects and perceptual granules that provide a basis for perceptual information systems useful in science and engineering. It is also

worthwhile to notice that there exists also methodology which incorporates Ziarko's idea of variable precision to DRSA methodology resulting in Variable Consistency Dominance Rough Set Approach (VCDRSA) [6].

There have been proposed also other modifications of RST, mainly changing the equivalence relation to weaker similarity relation [11], or defining equivalence relation in continuous attribute space without the need of discretization. Introduction of the structure onto the set of conditional attributes together with application of cluster analysis methodology for this purpose has been proposed by Cyran [1]. The applicability for the problem solved also with the use of CRSA in [3] has been demonstrated in the case study, but it is worth to say that the domain of possible use of modified indiscernibility relation extends to all problems with continuous attributes.

When considering modifications and improvements of Classical Rough Set Approach (CRSA) defined by Pawlak in [9] it may be of some interest to discuss the relation between given enhanced approach and the original CRSA. Basically there are two kinds of this relation: the first is when the modified approach is more general than the CRSA and then the CRSA is a special case of it, and the second is when the modified approach uses the inspiration from CRSA but in fact it defines a new methodology which cannot be reduced to CRSA.

The example of the first type is VPRSM, because CRSA is a special case of VPRSM with precision parameter set to one. Also the modified indisrenibility relation as defined by Cyran in [1] is more general than the original one, since the latter is a special case of the first. Contrary to these examples, the DRSA is such enhancement which cannot be reduced to classical rough sets: it is inspired by the notions present in RST, but the introduction of dominance relation for preference-ordered attributes (called criteria) instead of equivalence relation present in CRSA is the reason why CRSA cannot be derived from DRSA as its special case.

The DRSA is claimed to have many advantages over CRSA in applications with natural preference-ordered attributes. Not denying this statment in general, it is possible to demonstrate the example of such information system with preference-ordered attributes, which, when treated as a decision table, can yield better (in the sense of decision error) decision algorithm A than that generated by DRSA (A_{DRSA}). The superiority of A is also true (however in the sense of generality level) when the aforementioned algorithm A is compared with the algorithm A_{CRSA} obtained by application of CRSA. The quasi dominance rough set approach is the framework within which the algorithm A can be derived. That is why algorithm A will be referred to as A_{QDRSA}.

QDRSA can be considered as a hybrid of CRSA and DRSA. Like DRSA it is dedicated for problems with preference-ordered attributes, but contrary to DRSA, it does not resign from the classical indiscernibility. Therefore the relation I_{CRSA} and I_{QDRSA} are identical. It follows, that for the Information System $S = (U, Q, V_q, f)$ in which $Q = C \cup \{d\}$ and for any $x, y \in U$ the I_{QDRSA} is defined as

$$xI_{\mathrm{QDRSA}}y \Longleftrightarrow \forall q \in C f(x,q) = f(y,q). \tag{1}$$

The notions of lower and upper approximations, quality of approximation, (relative) cores, (relative) reducts and (relative) value reducts are therefore defined in QDRSA like in CRSA.

The consequence of this assumption is that QDRSA, like CRSA, requires discrete values of attributes. This is different from DRSA where corresponding notions rely on preference relation and this apprach does not require discrete attributes.

Similarly to DRSA (and contrary to CRSA), QDRSA is dedicated for problems with preference-ordered attributes, however, because it relies on (1) these attributes need to be of the discrete type. While in some problems it is a clear limitation, in others, namely in such which deal with attributes having inherently discrete nature, the use of classical indiscernibility relation (1) can be advantageous. The illustrative example, concerning real world application in evolutionary genetics, explaines this in more detail. Here, the second limitation of the QDRSA will be given. This limitation is the two-valued domain of the decision attribute $V_d = \{c0, c1\}$, where $c0 < c1$.

Certainly, aforementioned constraint excludes QDRSA from being applied in many problems having more complex decisions. However, there is a vast class of applications for which the binary decision is natural and sufficient. In such cases, if the preference-order is in adddition naturally assigned to the decision, the application of QDRSA can give better effects that either CRSA (which does not take into consideration the preference order) or DRSA (which resignes from indiscernibility relation, what, as it will be shown, can lead to suboptimal solutions).

In general, the types of decision rules obtained in QDRSA are identical to those generated by DRSA. However, because the decision attribute recognizes only two classes and due to relying on indiscernibility (instead of preference) relation, only two types (out of five possible in DRSA) are generated in QDRSA. These decision rules are of the types:

```
if q1 is at least v1 and
   q2 is at least v2 and
       ....          and
   qn is at least vn then
   decision is at least c1
```

and

```
if q1 is at most v1 and
   q2 is at most v2 and
       ....         and
   qn is at most vn then
   decision is at most c0
```

Certainly if only two classes are recognized the conlusions of the two above types of rules can be changed to *decision is c1* or *decision is c0* for the first and the second type respectively. However, for consistency with DRSA, the full syntax with phrases *at least* and *at most* will be used.

The conditions of the decision rules in QDRSA can be obtained from conditions of the correspnding rules in CRSA by introduction of the preference of attribute

values to these conditions. First, it requires the change of equalities to phrases like *at least* for the first type conclusion and *at most* for the second type conclusion. Second, it requires selection of minimal set of conditions (considering all decision rules for the given class), since for example the condition *q1 is at least 2* in one rule and *q1 is at least 3* in the other, are subject for dominance relation. This relation is crucial in DRSA: in QDRSA it is also important, but its realm is reduced to the final stage of the information retrival, as shown above. Therefore in QDRSA but not in DRSA the notion of relative value reduct is exploited with its full potential.

It is also worth to notice that not neccesserily the limitation of the types of decision rules to only two aforementioned is a drawback. For example, the lack of the fifth type of the decision rules possibly generated by DRSA is in fact a pure advantage in all problems with binary decision, since sensless in such conditions decision rules of the type

```
if ... then decision is at least c0 and at most c1
```

are never generated (contrary to DRSA which in certain situations can generate such rules).

In the subsequent text the syntax of QDRSA rules will be a little different. In this slightly modified syntax, the notation of the two types of rules available in QDRSA is more compact:

```
if q1 >= v1 and q2 >= v2 and .... and qn >= vn then at_least.C1
if q1 <= v1 and q2 <= v2 and .... and qn <= vn then at_most.C0
```

3 Illustrative Example

Human evolution at molecular level is reflected in the genome record. Some genes were under strong pressure of natural selection, while genetic variation in others is mainly the result of genetic drift and selectively neutral mutations. If the gene under consideration is exhibiting signatures of natural selection then some variants of it must be more or less fit to the environment. Very often it is associated with some disorder having genetic background, but in some cases it is responsible for the development of the species. The best known example of the latter is the ASPM gene responsible for the brain size in primates, including Humans [13].

There is also third type of selection in which heterozygotes (i.e., organisms having different alleles at two homologues chromosomes) are more fit than any homozygotes (i.e., organisms having identical variants at both homologues chromosomes). This is the case with human sickle cell anemia which is caused by two identical copies of mutated allele. However if this allele is present in heterozygote together with wild-type allele, then the carrier of one copy of mutant allele, not only does not suffer sickle cell anemia, but also is able to generate successful immune response to the malaria. Therefore, on malaria endemic regions the mutant allele is frequent, despite in homozygotes it is responsible for severe disorder.

The type of selection, described above, is called overdominance selection. It is one of the cases of balancing selection – the other case, called underdominance selection is proven to be unstable and the mutant alelle is relatively quickly eliminated

from the population. In the case of balancing selection caused by overdominance mechanism the mutant allele is kept in population for very long time, and sometimes it is even reflected by between-species polymorphism.

3.1 Neutrality Tests

Population geneticists have developed quite a number of statistical neutrality tests which serve to deny at given significance level the neutral Kimura's model. Positive signals generated by them can be interpreted as caused be the presence of natural selection. In the study we consider Tajima's T, Fu's D^* and F_s, Wall's Q and B, Kelly's Z_{nS} and Strobeck's S tests. The definition of them is beyond the scope of the paper, but they are summarized in [4]

When given gene is tested with the use of aforementioned tests, some of them can give positive while others negative signals. Moreover, positive signals can be caused by population expansion or gegraphical structure of the population. On the other hand the signatures of actual natural seletion can be supressed by the recombination. All these factors make the proper interpretation hard and not necesserily univocal.

Cyran and Kimmel have developed multinull hypotheses methodology (partially published in [4] and lately further improved) capable for the reliable interpretation of the test outcomes in the context of natural selection. However, since the method requires modified null hypotheses, the critical values of the tests are unknown and the huge amount of computer simulations must be carried out for estimation of these values.

Therefore, the author proposed application of artificial intelligence (AI) based methods for the interpretation based solely on the test outcomes against classical null hypotheses. In this methodology the battery of tests outcomes is considered as a set of conditional attributes and the expert knowledge is delivered from application of the multinull hypotheses for some small amount of genes. After crossvalidation of the model, the decision concerning other genes can be done based on testing only against classical null hypotheses and application of decision algorithm inferred with AI methodology.

As AI techniques, among others, the rough set approaches were applied. The comparison of CRSA with DRSA for this particular purpose is described in [2] where it is proved that neither CRSA nor DRSA generates decision algorithm which is optimal for the problem considered. The proof is done by a simple demonstration of another algorithm which is Pareto-preferred over both mentioned approaches. This algorithm can be obtained with QDRSA as presented below.

3.2 Decision Algorithms

Consider the information system $S = (U, Q, V_q, f)$ in which $Q = C \cup \{d\}$ The application of CRSA generates the following decision algorithm, referred here to as

$Algorithm_{\text{CRSA}}$ (Fig. 1). The outcomes of neutrality tests are designated as *NS*, *S*, and *SS* for non-significant, significant, and strongly significant, respectively.

```
BAL_SEL_DETECTED   = False
BAL_SEL_UNDETECTED = False
CONTRADICTION      = False
NO_DECISION        = False
if T = SS or (T = S and D* = S) or ZnS = S then
   BAL_SEL_DETECTED = True
if T = NS or (T = S and D* = NS and ZnS = NS) then
   BAL_SEL_UNDETECTED = True
if BAL_SEL_DETECTED and
   BAL_SEL_UNDETECTED) then
   CONTRADICTION = True
if not(BAL_SEL_DETECTED) and
   not(BAL_SEL_UNDETECTED) or
   CONTRADICTION then
   NO_DECISION = True
```

Fig. 1 $Algorithm_{\text{CRSA}}$, adopted from [2]

The algorithm generated by DRSA, called $Algorithm_{\text{DRSA}}$ is shown in Fig. 2.

```
at_least.BAL_SEL_DETECTED  = False
at_most.BAL_SEL_UNDETECTED = False
CONTRADICTION              = False
NO_DECISION                = False
if T >= SS or (T >= S and D* >= S) or ZnS >= S then
   at_least.BAL_SEL_DETECTED = True
if T <= NS or (T <= S and D* <= NS and ZnS <= NS) then
   at_most.BAL_SEL_UNDETECTED = True
if at_least.BAL_SEL_DETECTED and
   at_most.BAL_SEL_UNDETECTED then
   CONTRADICTION = True
if not(at_least.BAL_SEL_DETECTED)
   and not(at_most.BAL_SEL_UNDETECTED) or
   CONTRADICTION then
   NO_DECISION = True
```

Fig. 2 $Algorithm_{\text{DRSA}}$, adopted from [2]

It happened that the algorithm generated by QDRSA $Algorithm_{\text{QDRSA}}$ is identical to $Algorithm_{\text{DRSA}}$ when the whole universe U of the Information System S is used for algorithm genration. However, if the universe of the Information System S is

divided into two sets of rules: those used for information retrival in the process of generating the decision algorithm, and those left for testing, then the resulting algorithms generated by DRSA and QDRSA are different in some cases. Below we present only these algorithms which differ between the two approaches.

If the information about RECQL gene is excluded from the information system S and it is left for testing then the DRSA and QDRSA generate the algorithms $Algorithm_{\mathrm{DRSA}}(-RECQL)$ and $Algorithm_{\mathrm{QDRSA}}(-RECQL)$ respectively. Since the general structure of both algorithms is identical to that of $Algorithm_{\mathrm{DRSA}}$, only two crucial if-then rules (the ones after four initialization assignments, and before two contradiction/no-decision determining if-then rules) are presented in Figs. 3 and 4.

```
if (T >= S and D* >= S) or Zns >= S then
   at_least.BAL_SEL_DETECTED = True
if T <= NS or (D* <= NS and Zns <= NS) then
   at_most.BAL_SEL_UNDETECTED = True
```

Fig. 3 $Algorithm_{\mathrm{DRSA}}(-RECQL)$

```
if {T >= SS} or
   (T >= S and D* >= S) or Zns >= S  then
   at_least.BAL_SEL_DETECTED = True
if T <= NS or (D* <= NS and Zns <= NS) then
   at_most.BAL_SEL_UNDETECTED = True
```

Fig. 4 $Algorithm_{\mathrm{QDRSA}}(-RECQL)$

It is visible that the difference is in the existence of one more condition in the rule describing the detection of balancing selection. This condition reads 'if the outcome of Tajima test is at least strongly statistically significant'. It occurs in $Algorithm_{\mathrm{QDRSA}}(-RECQL)$ because condition $T = SS$ is a result of application of relative value reduct for one of the rules in the Information System $S(-RECQL)$ analyzed with CRSA. After changing it to $T >= SS$ when QDRSA is applied it is still not dominated by any other conditions detecting balancing selection. Since it is not dominated it must remain in the final decision algorithm presented above.

However, this is not the case in DRSA. This latter approach, when considering the dominance of decision rules for the class *at-least.BAL-SEL*, compares original (i.e., not reduced with relative value reduct) condition (A) $D^* >= S$ and $T >= SS$ and $Z_{nS} >= S$ with another original condition (B) $D^* >= S$ and $T >= S$ and $Z_{nS} >= NS$, instead of comparing (like QDRSA does) the condition (a) $T >= SS$ with condition (b) $D^* >= S$ and $T >= S$ being the results of application of relative value reducts in QDRSA sense to the original conditions (A) and (B), respectively.

It is clear that rule with condition (A) is dominated by rule with condition (B), and therefore condition (A) seemed to be redundant in DRSA sense for the

class *at-least.BAL-SEL*. However, rule with condition (a) is not dominated by rule with condition (b) and this is the reason why condition (a) is present in the $Algorithm_{\mathrm{QDRSA}}(-RECQL)$ while it is absent in $Algorithm_{\mathrm{DRSA}}(-RECQL)$. Conditions (B) and (b) in both approaches are necessary and reduced to condition (b) present in both algorithms.

Finally, let us consider what is the influence of inclusion of the condition $T >= SS$ to the $Algorithm_{\mathrm{QDRSA}}(-RECQL)$. When this algorithm is applied for the interpretation of neutrality tests for RECQL gene (i.e., the gene which was not present in the Information System $S(-RECQL)$ used for automatic information retrival) for four populations the decision error is reduced from 0.25 to 0. When the full jack-knife method of crossvalidation is applied, the decision error is reduced from 0.313 with DRSA, what seems rather unacceptable, to 0.125 with QDRSA. We have to mention that at the same time QDRSA *NO-DECISION* results have inreased from 0 to 0.188, however in the case of screening procedure for which this methodology is intended, the unsure decision is also an indication for more detailed study with the use of multi-null hypotheses methodology.

4 Discussion and Conclusions

DRSA is no doubt a powerfull tool for information retrieval from data representing preference ordered criteria. However, if the problem can be naturally reduced to discrete criteria and binary preference-ordered decision, then this sophisticated construction, designed to be as universal as possible, can be less efficient than QDRSA, dedicated for such type of applications.

The real world illustration is an example that such class of applications is of practical value, at least in all problems with automatic interpretation of a battery of statistical tests. The genetic example with neutrality tests is only one of them. Certainly, many other areas exist having similar properties from the information retrival point of view. In presented illustration the information preserved in the combination of neutrality tests has been retrieved by a novel method called QDRSA.

The comparison of QDRSA with CRSA gives the favor to the first when the preference-order is present in conditional and decision attributes. The resulting decision algorithms in QDRSA are more general, i.e they cover more points of the input space. Moreover, in many cases, because of possible domination of some QDRSA conditions over some other ones, the decision algoriths are shorter as comapred to CRSA. However, because the domination is checked after the application of relative value reducts, the negative effect (characteristic to DRSA) of ommiting the important condition from the decision algorithm (as it was shown in the illustrative example) is not present in QDRSA.

Acknowledgements. The scientific work was financed by Ministry of Science and Higher Education in Poland from funds for supporting science in 2008–2010, as a habilitation research project of the author. The project, registered under number N N519 319035, is entitled

'Artificial intelligence, branching processes, and coalescent methods in the studies concerning RNA-world and Humans evolution'.

References

1. Cyran, K.A.: Modified indiscernibility relation in the theory of rough sets with real-valued attributes: application to recognition of Fraunhofer diffraction patterns. In: Peters, J.F., Skowron, A., Rybiński, H. (eds.) Transactions on Rough Sets IX. LNCS, vol. 5390, pp. 14–34. Springer, Heidelberg (2008)
2. Cyran, K.A.: Classical and dominance based rough sets in the search for genes under balancing selection. In: Transactions on Rough Sets X. LNCS. Springer, Heidelberg (2009) (in press)
3. Cyran, K.A., Mrozek, A.: Rough sets in hybrid methods for pattern recognition. International Journal of Intelligent Systems 16(2), 149–168 (2001)
4. Cyran, K.A., Polanska, J., Kimmel, M.: Testing for signatures of natural selection at molecular genes level. Journal of Medical Informatics and Technologies 8, 31–39 (2004)
5. Greco, S., Matarazzo, B., Slowinski, R.: Rough approximation of preference relation by dominance relations. European Journal of Operational Research 117, 63–83 (1999)
6. Greco, S., Matarazzo, B., Slowinski, R., Stefanowski, J.: Variable consistency model of Dominance-based Rough Sets Approach. In: Ziarko, W.P., Yao, Y. (eds.) RSCTC 2000. LNCS, vol. 2005, pp. 170–181. Springer, Heidelberg (2001)
7. Kimura, M.: The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge (1983)
8. Pawlak, Z.: Rough sets. International Journal of Information and Computer Sciences 11, 341–356 (1982)
9. Pawlak, Z.: Rough Sets. Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dordrecht (1992)
10. Peters, J.F.: Near sets. general theory about nearness of objects. Applied Mathematical Sciences 1(53), 2609–2629 (2007)
11. Slowinski, R., Vanderpooten, D.: A generalized definition of rough approximations based on similarity. IEEE Transactions on Data and Knowledge Engineering 12(2), 331–336 (2000)
12. Zadeh, L.: Fuzzy sets. Information and Control 8(3), 338–353 (1965)
13. Zhang, J.: Evolution of the Human ASPM gene, a major determinant of brain size. Genetics 165, 2063–2070 (2003)
14. Ziarko, W.: Variable precision rough sets model. Journal of Computer and Systems Sciences 46(1), 39–59 (1993)