Rough Sets in the Interpretation of Statistical Tests Outcomes for Genes Under Hypothetical Balancing Selection

Krzysztof Cyran

Institute of Informatics, Silesian University of Technology,
Akademicka 16, 44-100 Gliwice, Poland
krzysztof.cyran@polsl.pl

Abstract. Detection of natural selection at the molecular level is one of the crucial problems in contemporary population genetics. There exists a number of statistical tests designed for it, however, the interpretation of the outcomes is often obscure, because of the existence of factors like population growth, migration and recombination. In his earlier work, the author has proposed the multi-null methodology, and he applied it for four genes implicated in human familial cancer: ATM, RECQL, WRN and BLM. Because of high computational effort required for estimating critical values under nonclassical nulls, mentioned methodology is not appropriate for selection screening. In the current paper, the author presents novel, rough set based methodology, helpful in the interpretation of tests outcomes applied versus only classical nulls. This method does not require long-lasting simulations and, as it is shown in the paper, it gives reliable results.

Keywords: rough sets, natural selection, ATM, BLM, RECQL, WRN, neutrality tests.

1 Introduction

Widely accepted Kimuras neutral model of evolution [1] states that, at the molecular level, the majority of genetic variation is caused by the selectively neutral forces like silent mutations and a genetic drift. Nevertheless, the model does not contradict the existence of selection at molecular level, although the role of it is not so important, as it had been thought before Kimuras work. When this work was published and, after some discussion, accepted, the majority of the genome was assumed to be selectively neutral. However, it is obvious that some mutations must be deleterious (and in fact we know many of such mutations causing serious genetic dysfunctions), some must be selectively positive (at least when the environment is changing) and some are known to be responsible for a phenomenon called balancing selection. Perhaps the most representative example of a positive selection is the ASMP locus, which is a major contributor to the brain size in primates [2,3]. Yet, even if the number of positive selections found grows

M. Kryszkiewicz et al. (Eds.): RSEISP 2007, LNAI 4585, pp. 716–725, 2007.
© Springer-Verlag Berlin Heidelberg 2007

up, the evidence for balancing selection is not so numerous. Therefore, the detection of the signatures for balancing selection operating at the molecular level remains one of the crucial problems in contemporary population genetics.

There exists a number of statistical non-neutrality tests [4,5,6,7] designed for the detection of such a selection in a gene under study. However, the interpretation of the outcomes of tests is hard because of the existence of factors like population growth, migration and recombination, which are not included in classical null hypothesis [8]. In his earlier work (published in part in [9] and in part unpublished), the author has proposed the multi-null hypothesis methodology, and using it, he was able to detect the signatures of a balancing selection in genes implicated in human familial cancer: in ATM (ataxia-telangiectasia mutated) and in a helicase involved in a repair of the DNA called RECQL. He also confirmed no evidence of such a selection in two other DNA helicases: WRN (Werners syndrome, see [10]) and BLM (Blooms syndrome, see [11]).

Because of high computational effort required for computing (by computer simulations) the critical values of the tests under nonclassical null hypotheses, the methodology proposed earlier is not appropriate as a screening tool. In a current paper the author presents rough set based methodology, helpful in the interpretation of tests outcomes, applied versus only classical nulls. The use of rough set theory for knowledge processing was dicated by the fact that test outcomes can be naturally discretized to a few values only, such as statistically non signinficant, statistically signinficant, or strongly statisctically significant. Moreover, since the critical values for classical null hypotheses are known, this method does not require time-consuming computer simulations and, as it is shown in the paper, it gives relatively reliable results.

2 Materials and Methods

As genetic material for this study, there was taken the single nucleotide polymorphisms (SNP) data, taken from the intronic regions of target genes. They form haplotypes, which can be used as tools to investigate the genetic diversity and possible disease associations. The first locus analyzed is ataxia-telangiectasia mutated (ATM) [12,13]. The ATM gene product is a member of a family of large proteins implicated in the regulation of the cell cycle and in the response to DNA damage. The other three genes include: human helicase RECQL, Blooms syndrome (BLM) and Werners syndrome (WRN). The products of these three genes are DNA helicases, enzymes involved in various types of DNA repair, including mismatch repair, nucleotide excision repair, and direct repair. A number of interesting facts about these genes were determined, including the question of selection signatures, addressed by the author and his co-workers [9].

The ATM gene, located in human chromosomal region 11q22-q23, spans 184 kb of genomic DNA. The intron-exon structure of the WRN locus spanning 186 kb at 8p12-p11.2 includes 35 exons, with the coding sequence beginning in the second exon. RECQL is composed of 15 exons, located at 12p12-p11 and spans 180 kb, whereas BLM, mapped to 15q26.1, has 22 exons and spans 154 kb. Blood samples for this study were collected from the residents of Houston, TX, from four major ethnic groups: Caucasians, Asians, Hispanics, and African-Americans.

To detect departures from the neutral model, the following statistics were used: Tajimas (1989) T (for uniformity, we follow here the nomenclature of Fu [5] and Wall [7]), Fu and Lis (1993) F^* and D^*, Kellys (1997) Z_{nS} and Walls (1999) Q and B, as well as Strobecks S test. The definitions of these statistics can be found in original works of the inventors, as well as, in a brief form, in Cyran et. al. (2004) pilot study [9].

In this study the rough set based method is used to simplify the process of determining whether the given gene is exhibiting the signatures of balancing selection or not. Such a selection (if present) is reflected by statistically significant departures from the null of the Tajimas and Fus tests towards positive values. However, not all such departures are indeed caused by a balancing selection [8], since such factors like population change in time, migration between subpopulations and a recombination can be reflected by similar outcomes of these tests. Therefore, a wide range of tests was included and the problem with the interpretation of their combinations occurred.

In order to apply a rough set based methodology, the decision table was built with tests outcomes treated as conditional attributes and a decision about the balancing selection treated as the only decision attribute. Fortunately, basing on previous studies, using multi-null methodology and heavy computer simulations, the author was able to determine the value of this decision attribute for given combination of conditional attributes. The purpose of this work was to propose and verify that the automatic and reliable interpretation of the battery of tests outcomes (perhaps without using all of them) can be done without application of the time consuming multi-null strategy. Therefore, to find the required set of tests, which is informative about the problem, there was applied the notion of a relative reduct with respect to decision attribute. Also, in order to obtain as simple decision rules as possible, the relative value reducts were used for particular elements of the Universe. To study the generalization properties and to estimate the decision error, the jack-knife crossvalidation technique was used.

3 Results and Discussion

The haplotypes for particular loci were inferred and their frequencies were estimated by using the Expectation-Maximization algorithm [14]. The results of tests T, D^*, F^*, S, Q, B and Z_{nS}, together with the decision concerning the evidence of balancing selection based on multi-null methodology, are given in Table 1.

The rough set based analysis of the Decision Table 1 reveals that there exist two relative reducts: $RED_1 = \{D^*, T, Z_{nS}\}$ and $RED_2 = \{D^*, T, F^*\}$. It is clearly visible that the core set is composed of tests D^* and T, whereas tests Z_{nS} and F^* can be chosen arbitrarily, according to the automatic data analysis. However, since it is known that both Fus tests F^* and D^* are examples of tests belonging to the same family, and therefore their outcomes are rather strongly correlated, it is advantageous to choose Kellys Z_{nS} instead of F^* test. It is so, because Z_{nS} outcomes are theoretically less correlated with outcomes of test D^*, belonging, as it was stated above, to the core and therefore required in any reduct. Generally, the same rule should be applicable also to the cases when the number of reducts is larger than two. However, the actual choice of the appropriate reduct in such a case can be more difficult, and the advise of a genetician should be of great relevance. The Decision Table 1 with set of conditional attributes reduced to the set RED_1 is presented in Table 2.

Table 1. The outcomes of the statistical tests for the classical null hypothesis. The table includes: Fus D^* test Walls B test, Walls Q test, Tajimas T test (known also as Tajimas D), Strobecks S test, Kellys Z_{nS} test, and Fus F^* test. The values of the test are: Non significant (NS) when $p > 0.05$, significant (*) if $0.01 < p < 0.05$, and strongly significant (**) when $p < 0.01$. The last column indicates the evidence or no evidence of balancing selection, based on the detailed analysis according to multi-null methodology.

		D^*	B	Q	T	S	Z_{nS}	F^*	Balancing selection
ATM	AfAm	*	NS	NS	*	NS	NS	*	Yes
	Cauc	*	NS	NS	**	**	*	**	Yes
	Asian	NS	NS	NS	*	NS	*	NS	Yes
	Hispanic	*	NS	NS	**	NS	*	*	Yes
RECQL	AfAm	NS	NS	NS	**	NS	NS	NS	Yes
	Cauc	*	NS	NS	**	NS	NS	**	Yes
	Asian	NS	*	*	*	NS	*	NS	Yes
	Hispanic	*	NS	NS	**	NS	NS	*	Yes
WRN	AfAm	NS	NS	NS	NS	NS	NS	NS	No
	Cauc	*	NS	NS	NS	NS	NS	NS	No
	Asian	*	NS	NS	NS	NS	NS	NS	No
	Hispanic	NS	NS	NS	NS	NS	NS	NS	No
BLM	AfAm	NS	NS	NS	NS	NS	NS	NS	No
	Cauc	NS	NS	NS	*	NS	NS	*	No
	Asian	NS	NS	NS	NS	NS	NS	NS	No
	Hispanic	NS	NS	NS	NS	NS	NS	NS	No

After a reduction of the set of informative tests to set $RED_1 = \{D^*, T, Z_{nS}\}$, there was considered the problem of coverage of the discrete space generated by these statistics, by the examples included in the training set. The results are given in Table 3, and they reveal that in such a space the fraction of points, which are included in training data, is only 30%. The next step was to apply the notion of the relative value reducts to particular decision rules in the Decision Table 2. The resulting new Decision Table is presented in Table 4. Basing on this table, the Decision Algorithm 1 was obtained. Note that this algorithm is simplified as compared to the algorithm that corresponds to the Decision

Table 2. The Decision Table, in which the set of tests is reduced to relative reduct RED_1 composed of tests: D^*, T, and Z_{nS}

		D^*	T	Z_{nS}	Balancing selection
ATM	AfAm	*	*	NS	Yes
	Cauc	*	**	*	Yes
	Asian	NS	*	*	Yes
	Hispanic	*	**	*	Yes
RECQL	AfAm	NS	**	NS	Yes
	Cauc	*	**	NS	Yes
	Asian	NS	*	*	Yes
	Hispanic	*	**	NS	Yes
WRN	AfAm	NS	NS	NS	No
	Cauc	*	NS	NS	No
	Asian	*	NS	NS	No
	Hispanic	NS	NS	NS	No
BLM	AfAm	NS	NS	NS	No
	Cauc	NS	*	NS	No
	Asian	NS	NS	NS	No
	Hispanic	NS	NS	NS	No

Table 2. At the same time, it is more general, which can be observed in Table 5, presenting the information analogous to Table 3. In Table 5, the coverage of points is based on the number of points which are classified with the use of the simplified Algorithm 1. One should notice that the fraction of points covered by algorithm is 74%, however, since 11% is classified as both with and without the evidence of balancing selection, therefore only 63% of the points could be treated as really covered.

Algorithm 1

```
BALANCING_SELECTION If: T = ** or (T = * and D* = *) or ZnS = *
NO_SELECTION If: T = NS or (T = * and D* = NS and ZnS = NS)
```

Purely automatic knowledge processing technique resulting in Algorithm 1, can be further improved by supplying it with the additional information, concerning the domain under study. It is clearly true that, if a balancing selection is determined by the statistical significance of the given test, then such a selection is even more probable when the outcome of this test is strongly statistically significant.

Therefore, instead of equalities in Algorithm 1, there are proposed inequalities in the generalized version referred to as Algorithm 2. Such inequality means that the given test is at least of the value of statistical significance shown to the right of the inequality sign, but it can obviously be also more significant. In other words, the main difference between Algorithm 2, as compared to the Algorithm

Table 3. The discrete space of three tests: D^*, T and Z_{nS}. The domain of each test outcome (coordinate) is composed of three values: ** (strong statistical significance $p < 0.01$), * (statistical significance $0.01 < p < 0.05$), and NS (no significance $p > 0.05$). The given point in a space is assigned to: S (the evidence of balancing selection), N (no evidence of balancing selection) or empty cell (point not covered by the training data). The assignment is done basing on raw training data with conditional part reduced to the relative reduct RED_1 . Note that the fraction of points covered by training examples is only 30%.

		T								
		$**$			$**$			NS		
		Z_{nS}			Z_{nS}			Z_{nS}		
		$**$	$*$	NS	$**$	$*$	NS	$**$	$*$	NS
	$**$									
D^*	$*$		S	S			S			N
	NS			S		S	N			N

Table 4. The set of tests is reduced to reflect the relative reduct composed of tests: D^*, T, and Z_{nS}, and additionally the notion of relative value reduct is used to further reduce the complexity of separate rows in a decision table

		D^*	T	Z_{nS}	Balancing selection
ATM	AfAm	*	*		Yes
	Cauc		**		Yes
	Asian			*	Yes
	Hispanic		**		Yes
RECQL	AfAm		**		Yes
	Cauc		**		Yes
	Asian			*	Yes
	Hispanic		**		Yes
WRN	AfAm		NS		No
	Cauc		NS		No
	Asian		NS		No
	Hispanic		NS		No
BLM	AfAm		NS		No
	Cauc	NS	*	NS	No
	Asian		NS		No
	Hispanic		NS		No

1, is that instead of formulas of the type $testoutcome = *$ it uses formulas of the type $testoutcome >= *$, meaning that the test outcome is at least significant (and perhaps strongly significant).

Algorithm 2 deals also with the problem of contradiction, and in such a case, it generates no decision about the evidence of balancing selection in a gene under study. The problem of covering points in a discrete space generated by three

Table 5. The discrete space of three tests: $D*$, T and Z_{nS}, forming a relative reduct. The domain of each test outcome (coordinate) is composed of three values: ** (strong statistical significance $p < 0.01$), * (statistical significance $0.01 < p < 0.05$), and NS (no significance $p > 0.05$). The given point in a space is assigned to S and N (with the meaning identical to that given in the caption of Table 3), or "-" having the meaning of contradiction between evidence and no evidence of the balancing selection. The space is filled basing on the simplified Decision Algorithm 1, which uses the relative value reducts varying among different training examples. Note that the fraction of points covered is now 74%, but it includes 11% denoting the contradicting decisions, and such a case should be treated as the lack of decision. Therefore, the real fraction of points assigned with some decision is now 63%.

		T								
		$**$			$**$			NS		
		Z_{nS}			Z_{nS}			Z_{nS}		
		$**$	$*$	NS	$**$	$*$	NS	$**$	$*$	NS
D^*	$**$	S	S	S					-	N
	$*$	S	S	S	S	S	S		-	N
	NS	S	S	S		S	N		-	N

tests in Algorithm 2 is presented in Table 6. This table shows that all points are covered by Algorithm 2, yet since 22% are designated as contradictions, therefore 78% points in a space are really recognizable by this algorithm.

Moreover, the remainding fraction of 22% of points with no decision assigned to them, are such points which denote situations that are extremely rare from genetics point of view. Namely, these are the situations where the outcome of the Tajima test T is non significant and, at the same time, the outcome of the Kelly Z_{nS} test is significant or even strongly significant. Such a situation has never happened for any gene, for any population and for any of the null hypothesis, considered in the detailed multi-null study. Therefore, even if one cannot totally exclude such situations from theoretical point of view, in practice one meets them very rarely.

Algorithm 2

```
BALANCING_SELECTION := False; NO_DECISION := False;
If T >= ** or (T >= * and D* >= *) or ZnS >= * then
   BALANCING_SELECTION := True;
If T = NS or (T = * and D* = NS and ZnS = NS) then
   If BALANCING_SELECTION then
      NO_DECISION := True
   else
      BALANCING_SELECTION := False;
```

The comparison of Table 3 with Tables 5 and 6 shows the degree of generalization (into unknown combinations of the tests outcomes). It was increased by the application of rough set theory (Table 5) and by additional genetic knowledge (Table 6). Both these strategies, when applied together, resulted in a relatively

Table 6. The discrete space of three tests: D^*, T and Z_{nS}, forming a relative reduct. The domain of each test outcome (coordinate) is composed of three values: ** (strong statistical significance $p < 0.01$), * (statistical significance $0.01 < p < 0.05$), and NS (no significance $p > 0.05$). The given point in a space is assigned to S, N or "-" (with the meaning identical to that given in captions of Tables 3 and **??**). If any character is in parentheses, it means, that the point is assigned to the given value not automatically. Rather, the simple reasoning is used. It states that the selection is even more probable for given test showing strong significance (**), when automatic knowledge acquisition indicated such selection for this test being just significant (*) with the values of other tests unchanged. The assignment in Table 6 is done basing on the Decision Algorithm 2, which, similarly to Algorithm 1, uses the relative value reducts varying among different training examples. Note that the fraction of points covered by the algorithm is now 100%, but 22% denotes the contradiction in the decision, and such a case should be treated as the lack of decision. Therefore, the fraction of points really assigned with the decision is now 78%.

					T					
			**			**			NS	
			Z_{nS}			Z_{nS}			Z_{nS}	
		**	*	NS	**	*	NS	**	*	NS
	**	S	S	S	(S)	(S)	(S)	(-)	-	N
D^*	*	S	S	S	S	S	S	(-)	-	N
	NS	S	S	S	(S)	S	N	(-)	-	N

high increase of covering of the space generated by test outcomes (from 30% covered by the training examples, to 78% covered by the Algorithm 2).

However, here the question could be raised, what is the probability of correct generalization into unknown situations. To study this problem, there was applied automatic knowledge extraction procedure presented above, in the so-called jack-knife cross-validation, which is known to be unbiased in estimating the decision error of any classifier. Classical jack-knife strategy assumes that the training is performed basing on all-but-one training examples, and that the testing is done for the excluded example. After iterating this procedure N times (where N is the number of training facts), the average of decision errors in separate iterations is an unbiased estimate of the decision error. However, in case considered such a strategy could give too optimistic results, because training facts describing one gene in four different populations are not independent, and even after excluding one of them some knowledge about it is passed to the classifier. That is why, to be rigorous about the conclusions, the author decided to exclude from the iterations all four examples concerning one particular gene, and perform training basing on examples concerning three remaining genes.

The detailed presentation of results of cross-validation is beyond the scope of this paper. Here, the author would only like to point out that relatively large decrease of the number of training examples, which was the result of the assumed strategy, could produce pessimistic estimates of the decision error. However, it proved that even such pessimistic estimate as can be seen in Table 7, is small enough (12.5% with a variation between iterations equal to 0.0156) to claim that the proposed methodology could be utilized as useful tool in looking for candidates for more detailed analysis with computationally more requiring strategy, like the multi-null methodology. The last statement is based on the fact that as much as 87.5% correct recognitions of balancing selection for unknown genes were done when the proposed rough set based methodology was applied, with completely no need for performing long-lasting computer simulations for calculation of critical values of tests under non-classical null hypotheses (as required by multi-null methodology). The results of cross-validation procedure are also summarized in a form of confusion matrix in Table 8.

Table 7. The results of the cross-validation in a modified jack-knife strategy

Iteration	Errors in populations				Percentage of	Decision
without gene	African-American	Caucasian	Asian	Hispanic	correct decisions	Error
ATM	*Y*	*N*	*N*	*N*	75%	25%
RECQL	*N*	*N*	*N*	*N*	100%	0%
WRN	*N*	*N*	*N*	*N*	100%	0%
BLM	*N*	*Y*	*N*	*N*	75%	25%
Average:					**87.5%**	**12.5%**

Table 8. The confussion matrix of the cross-validation test

		Prediction	
		Lack of balancing selection	Balancing selection
Actual value	Lack of balancing selection	7	1
	Balancing selection	1	7

4 Conclusion

Since the time of Kimura's book [1] the search for the signatures of natural selection at molecular level has become one of important directions in genetics. However, many non-neutrality tests generate similar patterns for such depatures from neutral model like population growth or substructure of population. Since these factors influence different tests in a different way, the battery of tests can be more informative than any separate one. The problem of interpretation of a battery of such tests was considered in a paper. It proved that the rough set based decision making system can correctly (i.e with the concordance with time consuming mulit-null methodology) recognize 87.5% of cases of balancing selection for genes not used in a training.

Acknowledgments. The author would like to acknowledge the financial support of this work in the genetics related part under the MNiSW grant number

3T11F 010 29, and in the rough set related part under author's activities BW2007 at SUT. The author would like also to thank to Prof. M. Kimmel from Rice University in Houston for long discussions and advise concerning the multi-null methodology using statistical non-neutrality tests.

References

1. Kimura, M.: The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge (1983)
2. Zhang, J.: Evolution of the Human ASPM Gene, a Major Determinant of Brain Size. Genetics 165, 2063–2070 (2003)
3. Evans, P.D., Anderson, J.R., Vallender, E.J., Gilbert, S.L., Malcom, Ch.M. et al.: Adaptive Evolution of ASPM, a Major Determinant of Cerebral Cortical Size in Humans. Human Molecular Genetics 13, 489–494 (2004)
4. Fu, Y.X., Li, W.H.: Statistical Tests of Neutrality of Mutations. Genetics 133, 693–709 (1993)
5. Fu, Y.X.: Statistical Tests of Neutrality of Mutations Against Population Growth, Hitchhiking and Background Selection. Genetics 147, 915–925 (1997)
6. Kelly, J.K.: A Test of Neutrality Based on Interlocus Associations. Genetics 146, 1197–1206 (1997)
7. Wall, J.D.: Recombination and the Power of Statistical Tests of Neutrality. Genet. Res. 74, 65–79 (1999)
8. Nielsen, R.: Statistical Tests of Selective Neutrality in the Age of Genomics. Heredity 86, 641–647 (2001)
9. Cyran, K.A., Polaska, J., Kimmel, M.: Testing for Signatures of Natural Selection at Molecular Genes Level. J. Med. Inf. Techn. 8, 31–39 (2004)
10. Dhillon, K.K., Sidorova, J., Saintigny, Y., Poot, M., Gollahon, K., Rabinovitch, P.S., Mon-nat Jr., R.J.: Functional Role of the Werner Syndrome RecQ Helicase in Human Fibroblasts. Aging Cell 6, 53–61 (2007)
11. Karmakar, P., Seki, M., Kanamori, M., Hashiguchi, K., Ohtsuki, M., Murata, E., Inoue, E., Tada, S., Lan, L., Yasui, A., Enomoto, T.: BLM is an Early Responder to DNA Double-strand Breaks. Biochem. Biophys. Res. Commun. 348, 62–69 (2006)
12. Golding, S.E., Rosenberg, E., Neill, S., Dent, P., Povirk, L.F., Valerie, K.: Extracellular Signal-Related Kinase Positively Regulates Ataxia Telangiectasia Mutated, Homologous Recombination Repair, and the DNA Damage Response. Cancer Res. 67, 1046–1053 (2007)
13. Schneider, J., Philipp, M., Yamini, P., Dork, T., Woitowitz, H.J.: ATM Gene Mutations in Former Uranium Miners of SDAG Wismut: a Pilot Study. Oncol. Rep. 17, 477–482 (2007)
14. Polanska, J.: The EM Algorithm and its Implementation for the Estimation of the Frequencies of SNP-Haplotypes. Int. J. Appl. Math. Comp. Sci. 13, 419–429 (2003)