Biocybernetics and Biomedical Engineering
2008, Volume 28, Number 3, pp. 45–50

Application of the *k*-NN Classifier for Mutagenesis Tests. Recognition of Wild Type and Defective in DNA Repair Bacterial Strains on the Basis of Adaptive Response to Alkylating Agents

AGNIESZKA M. MACIEJEWSKA[1], ADAM JÓŹWIK[2,*], JAROSŁAW T. KUŚMIEREK[1], BEATA SOKOŁOWSKA[3]

[1] *Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warsaw, Poland*
[2] *Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Sciences, Warsaw, Poland*
[3] *Medical Research Center, Polish Academy of Sciences, Warsaw, Poland*

The *k*-Nearest Neighbor classifier (*k*-NN) was applied to differentiate two bacterial strains, the wild type and its mug derivative. Bacterial cells were exposed to different concentrations of chloroacetaldehyde and studied under two different conditions, i.e. with and without induction of an adaptive response. We evaluated the influence of adaptation on the wt and mug strains by estimating the probability of misclassification to the class: *no adaptation* or *adaptation.* We have also checked differentiation between *wt* and *mug*, separately for *adapted* and *non-adapted* conditions. Our results confirm the usefulness of the *k*-NN classifier as a tool for statistical analysis of results of mutagenesis tests.

K e y w o r d s: pattern recognition, *k*-NN classifier, DNA repair, adaptive response, mutagenesis, mug

1. Introduction

Alkylating agents are environmental genotoxic agents with mutagenic and carcinogenic potential. One of them is 2-chloroacetaldehyde (CAA), a metabolite of vinyl chloride. CAA causes modifications in DNA which include ethenoadducts, such as 3,N^4-ethenocytosine, 1,N^6-ethenoadenine, N^2,3-ethenoguanine and 1,N^2-ethenoguanine. These exocyclic DNA adducts are also generated endogenously by products of peroxidation of membrane lipids [1].

* Correspondence to: Adam Jóźwik, Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Sciences, ul. Ks. Trojdena 4, 02-109 Warsaw, e-mail: ajozwik@ibib.waw.pl
Received 21 February 2008; Accepted 05 May 2008

Cells are well equipped with DNA repair mechanisms which protect them against harmful effect of alkylating agents. Among the repair mechanisms there is adaptive response (Ada response) involving *ada*, *alkB*, *alkA* and *aidB* genes expressed after exposure to non-toxic doses of direct-acting alkylating agents. These inducible genes encode proteins responsible for repair of DNA damages, including damages caused by CAA [2]. It is known that AlkA glycosylase removes N^2,3-ethenoguanine [3] whereas 3,N^4-ethenocytosine and 1,N^6-ethenoadenine are substrates for AlkB oxygenase [4, 5]. Mismatch uracil glycosylase (Mug) is not involved in the Ada response system, and its primary function is the repair of 3,N^4-ethenocytosine. It is also able to remove 1,N^2-ethenoguanine from DNA [6,7].

Algorithms of pattern recognition are often applied in biomedical studies [8–12]. One of the most popular and highly effective unsupervised classification techniques is based on the *k*-Nearest Neighbor (*k*-NN) rule. Here, we present the usefulness of the *k*-NN rule in differentiation between two bacterial strains, the wild type (*wt*) and its derivative, *mug*. The *mug* strain carries a disrupted version of mismatch uracil glycosylase gene. It is defective in removing 3,N^4-ethenocytosine and 1,N^2-ethenoguanine from DNA and hence cumulates mutations caused by these ethenoadducts. Induction of Ada response decreases the mutagenic effect in the studied strains.

2. Materials and Methods

2.1. Experimental

Details of the biological methods used are described in [13]. Briefly, we have modified IF102 plasmid DNA *in vitro* with 50, 100 and 200 mM CAA to generate ethenoadducts. A control plasmid was treated in the same way as the modified one, except that CAA was absent from the incubation mixture. These plasmids were introduced into bacterial cells *in vivo* and then mutants were selected. AM1-wild type (*wt*) and AM3 (*mug*) bacterial strains were used. Both of them were studied under two different conditions, with and without induction of the adaptive response, denoted as *A* and *NA*, respectively. Finally, the effect caused by IF 102 DNA damages (ethenoadducts), described here as the mutagenesis level, was analyzed.

2.2. Pattern Recognition Method Based on the k-NN Rule

One of the most popular and highly effective methods of statistical pattern recognition is based on the *k* Nearest Neighbor rule (*k*-NN rule) [14–16]. The standard *k*-NN classifier assigns the classified object to the same class as the majority of its *k* nearest neighbors (i.e., nearest points) in the reference set. The value of *k* is established experimentally. A simple two-dimensional example given in Fig. 1 illustrates how the *k*-NN rule operates.

The value of k should be determined in a way that offers the smallest probability of misclassification. Such a k is called the optimum one and the k-NN is then called the optimum k-NN rule. The probability of misclassification can be estimated by the *leave-one-out* method [14] on basis of the reference set R. Each point $\boldsymbol{x}$ from the reference set is classified, using the k-NN rule, to the class most heavily represented among its k nearest neighbors found in the set R-$\{\boldsymbol{x}\}$. The probability of misclassification is estimated by an error rate, $E_r(k)=r/m$, where r is the number of misclassified objects and m is the number of classified points, i.e., the number of points in the set R.

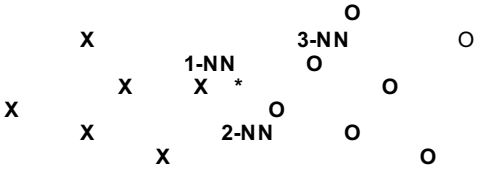

Fig. 1. An illustration of the k-NN rule.The symbols 1-NN, 2-NN and 3-NN denote the first, the second and the third nearest neighbor respectively. The point „*" is qualified to class 2 since two out of its three nearest neighbors come from this class ("X" – points from class 1, "O" – points belonging to class 2)

Figure 2 illustrates the application of the *leave-one-out* method to the small one-dimensional reference set.

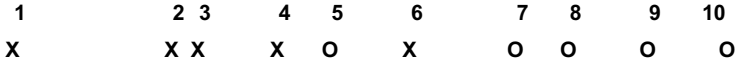

Fig. 2. An example illustrating the leave one out method. The 1-NN rule misclassifies three points, 4, 5 and 6 whereas the 3-NN rule misclassifies two points, 5 and 6 ("X" – points from class 1, "O" – points belonging to class 2)

In the example given in the Fig. 2, the error rate $E_r(1)$ for the 1-NN rule equals 0.3 since three points (4, 5 and 6) out of ten points are misclassified. The nearest neighbors of these three points come from the opposite class. We can check that in case of the 3-NN rule the error rate E_r=0.2, because the number of misclassified points decreased to two (points 5 and 6). Two out of the three nearest neighbors of point 4 (i.e. the majority) come from class 1, i.e., from the same class to which point 4 belongs. Thus, this time point 4 is correctly classified. Hence, the error rate $E_r(3)$ for the 3-NN rule is lower than $E_r(1)$.

To find the optimum value of k, it is necessary to calculate the error rates $E_r(k)$ for all possible values of k, i.e., for k=1, 2,…, m–1, and select the k that corresponds to the lowest value of the error rate $E_r(k)$.

The *leave-one-out* method is more economical than the testing set approach since we do not need to divide our data set into the reference set and the testing set. Hence, using the *leave-one-out* method we evaluate the k-NN classifier based on the whole data set.

2.3. Data Set and Analysis

The data set consisted of mutagenesis level values obtained from results (referred to as objects) of twelve independent experiments on each strain and for each CAA concentration (see Experimental). In summary, 96 objects were analyzed, 24 points for each strain and for each adaptation condition. The analysis with the use of the *k*-NN rule was performed by two biological approaches. The first approach involves distinguishing the conditions: without adaptation (no adaptation, symbol *NA* – as class 1) and with induction of the adaptive response (with adaptation, symbol *A* – as class 2) independently for each bacterial strain. The second approach involves distinguishing both bacterial strains (the *wt* strain – as class 1, the *mug* one – as class 2) independently with and without adaptation.

3. Results and Discussion

The *mug* strain is defective in repair of ethenoadducts to DNA, and hence its mutagenesis level after exposure to CAA is higher than in case of the *wt*. Induction of the Ada-response decreases the mutagenesis level by induction of DNA repair enzymes involved in removal of ethenoadducts. These are well known facts confirmed by us and others [13, 17–19].

In this work we present the *k*-NN classifier as a statistical tool in mutagenesis tests. The experimental data used here for illustrating the method were taken from a larger project [13] and subjected to *k*-NN classifier analysis. The results of the analysis are presented in Tables 1 and 2. The lower the misclassification rate (E_r), the greater is the difference between the considered classes. To evaluate the influence of adaptation on *wt* and *mug* strains, the probability of misclassification of the class of *no adaptation* and with *adaptation* was estimated (Table 1). The E_r values, were also calculated when differentiation between *wt* and *mug* was investigated, separately for *A* and *NA* conditions (Table 2).

Misclassification rates were very high in control conditions (values >30%), regardless of which of the two pairs of classes were differentiated (the first row of

Table 1. Misclassification rates (E_r). Recognition between no adaptation (NA – class 1) and with adaptation (A – class 2) for the wild type (*wt*) and its derivative (*mug*) strain

Concentration of CAA [mM]	*NA vs. A*, E_r [%]	
	_wt	*_mug*
0 (control)	41.7	45.8
50	29.2	0.0
100	0.0	0.0
200	0.0	8.3

Table 2. Misclassification rates (E_r). Recognition between the wild type (class 1) and the mug strain (class 2) under condition without (*NA*) and with adaptation (*A*)

Concentration of CAA [mM]	*_wt vs. mug*, E_r [%]	
	NA	*A*
0 (control)	50.0	33.3
50	0.0	8.3
100	0.0	4.2
200	8.3	0.0

results in Tables 1 and 2). This indicates that there is no difference in mutagenesis level (and in DNA damage repair) between *wt* and *mug* strains when bacterial cells are not exposed to mutagens and induction of Ada response has no effect. However, for concentrations of 50 mM CAA the error rate is equal to zero and 29.2% in case of *mug* and *wt* strain, respectively. The classifier quite correctly recognizes the influence of induction of the Ada response in experiments with the *wt* strain at 100 and 200 mM CAA. However, perfect recognition is only at 50 and 100 mM CAA in the *mug* strain. At 200 mM CAA the error rate is 8.3% (Table 1). These misclassification rates show that the *mug* strain is more sensitive to CAA than the *wt*. They indicate also that 100 mM CAA concentration is the best for observation of adaptation effect in this experiment, whereas 50 mM CAA concentration is to a low (and there is no effect for the *wt*) and 200 mM is to a high (and the error in the experimental results may be a large). All these findings are in agreement with those previously described.

One the other hand, the recognition between *wt* and *mug* in adapted strains is perfect only at 200 mM CAA although misclassification rates decrease with an increase in CAA concentration. At lower concentrations of CAA the mutagenic effect of the lack of Mug glycosylase in the *mug* strain may be masked by the repair activity of enzymes induced by the Ada response. Without those repair enzymes (*NA*) both strains can be correctly recognized at 50 and 100 mM CAA, but at 200 mM CAA the misclassification rate increases slightly (Table 2).

The results of this analysis confirm that *k*-NN is a very useful tool in our experimental model. We believe that the classifier may serve to differentiate between bacterial strains exposed to mutagenic agents. In fact, it may help to interpret results of biological sets of data performed for evaluation of different aspects of strain behaviour in bacterial models.

Acknowledgments

We wish to thank Ms Anna Karcz for her assistance in some experiments. This work was supported by the Ministry of Science and Higher Education, Poland, Grant N301 065 31/1979.

References

1. Bartsch H.: Exocyclic adducts as new risk markers for DNA damage in man. In: Exocyclic DNA Adducts in Mutagenesis and Carcinogenesis. Singer B., Bartsch H, (Eds), IARC Scientific Publication No. 150, Lyon 1999, 1–16.
2. Sedgwick B., Lindahl T.: Recent progress on the Ada response for inducible repair of DNA alkylation damage. Oncogene. 2002, 21, 8886–8894.
3. Matijasevic Z., Sekiguchi M., Ludlum DB.: Release of N2,3-ethenoguanine from chloroacetaldehyde--treated DNA by Escherichia coli 3-methyladenine DNA glycosylase II. Proc. Natl. Acad. Sci. USA 1992, 89(19), 9331–9334.
4. Delaney J.C., Smeester L., Wong C., Frick L.E., Taghizadeh K., Wishnok J.S., Drennan C.L., Samson L.D., Essigmann J.M.: AlkB reverses etheno DNA lesions caused by lipid oxidation in vitro and in vivo. Nat. Struct. Mol. Biol. 2005, 12, 855–860.
5. Mishina Y., Yang C.G., He C.: Direct repair of the exocyclic DNA adduct 1,N^6-ethenoadenine by the DNA repair AlkB proteins. J. Am. Chem. Soc. 2005, 127, 14594–14595.
6. Saparbaev M., Laval J.: 3,N^4-ethenocytosine, a highly mutagenic adduct, is a primary substrate for Escherichia coli double-stranded uracil-DNA glycosylase and human mismatch-specific thymine--DNA-glycosylase. Proc. Natl. Acad. Sci. USA 1998, 95, 8508–8513.
7. Saparbaev M., Langouët S., Privezentzev C.V., Guengerich F.P., Cai H., Elder R.H., Laval J.: N(2)-ethenoguanine, a mutagenic DNA adduct, is a primary substrate of Escherichia coli mismatch--specific uracil-DNA glycosylase and human alkylpurine-DNA-N-glycosylase. J. Biol. Chem. 2002, 277(30), 26987–26993.
8. Saastamoinen A., Huupponen E., Varri A., Hasan J., Himanen S-L.: Computer program for automated sleep depth estimation. Comp. Meth. Program. Biomed. 2006, 82, 58–66.
9. Shi H., Paolucci U., Vigneau-Callahan K.E., Milbury P.E., Matson W.R., Kristal B.S.: Development of biomarkers based on diet-dependent metabolic serotypes: practical issues in development of expert systems-based classification models in metabolomic studies. OMICS J. Int. Biol. 2004, 8(3), 197–208.
10. Sokołowska B., Jóźwik A.: Recognition of cycles of repeated hypoxia on the basis of time periods in biological model. In: Kurzyński M., Puchala E., Wozniak M., Zolnierek A., Computer Recognition Systems2, ASC 45, Springer-Verlag Berlin Heidelberg 2007, 778–785.
11. Sokołowska B., Jóźwik A., Pokorski M.: A fuzzy-classifier system to distinguish respiratory pattern evolving after diaphragm paralysis in the cat. Jpn. J. Physiol. 2003, 53(4), 301–307.
12. Tsai C-A., Chen C-H., Lee T-C., Ho I-C., Yang U-C., Chen JJ.: Gene selection for sample classifications in microarray experiments. DNA and Cell Biol. 2004, 23(10), 607–614.
13. Maciejewska A. et al., 2008, in preparation.
14. Devijver P.A., Kittler J.: Pattern Recognition: A Statistical Approach, Prentice Hall, London 1982.
15. Duda R.O., Hart P.E., Stock D.G.: Pattern Classification, John Wiley and Sons, New York 2001.
16. Fix E., Hodges J.L.: Discriminatory Analysis. Nonparametric Discrimination Small Sample Performance, project 21-49-004, Report Number 11, USAF School of Aviation Medicine, Randolph Field, Texas 1952, 280-322, reprinted in the book: Dasarathy B.V. “NN Pattern Classification Techniques”, IEEE Computer Society Press 1991, 40–56.
17. Borys E., Mroczkowska-Slupska M.M., Kusmierek J.T.: The induction of adaptive response to alkylating agents in Escherichia coli reduces the frequency of specific C→T mutations in chloroacetaldehyde-treated M13 glyU phage. Mutagenesis 1994, 9, 407–410.
18. Jurado J., Maciejewska A., Krwawicz J., Laval J., Saparbaev M.K.: Role of mismatch-specific uracil--DNA glycosylase in repair of 3,N^4-ethenocytosine in vivo. DNA Repair. 2004, 3, 1579–1590.
19. Mroczkowska M.M., Kolasa I.K., Kusmierek J.T.: Chloroacetaldehyde-induced mutagenesis in Escherichia coli: specificity of mutations and modulation by induction of the adaptive response to alkylating agents. Mutagenesis 1993, 8, 341–348.