Nucleic Acids Research, 2007, Vol. 35, Web Server issue ***W683–W687***
doi:10.1093/nar/gkm262

SePaCS—a web-based application for classification of seroreactivity profiles

Andreas Keller[1,*], Nicole Comtesse[2], Nicole Ludwig[2], Eckart Meese[2] and Hans-Peter Lenhof[1]

[1]Center for Bioinformatics, Saarland University, Building E1 1, 66041 Saarbrücken, Germany and [2]Department of Human Genetics, Medical School, Saarland University, Building 60, 66421 Homburg/Saar, Germany

Received January 31, 2007; Revised March 29, 2007; Accepted April 8, 2007

ABSTRACT

Immunogenic antigen sets possess high potential for minimally invasive disease detection and monitoring. For various diseases, including cancer, appropriate antigen sets have already been detected in blood sera of patients. Typically, a large number of sera from diseased and unaffected persons is screened for the antigens of interest. Sophisticated statistical learning approaches are trained on the resulting data set to classify sera as either tumor or normal sera. We developed a web-based application, called 'Seroreactivity Profile Classification Service' (SePaCS) that enables clinical groups to carry out analyzes of training sets and predictions of unclassified seroreactivity profiles with minimal effort. SePaCS provides a broad range of classification methods: four versions of a Naïve Bayes Classifier, Support Vector Machines with a radial basis function kernel, Linear Discriminant Analysis, and Diagonal Discriminant Analysis. The computed results are summarized in a PDF file. We demonstrate the functionality of SePaCS exemplarily for meningioma, a generally benign intracranial tumor. As a second example, we evaluated SePaCS on glioma, a malignant brain tumor. SePaCS is freely available at http://www.bioinf.uni-sb.de/sepacs.

INTRODUCTION

Tumor markers are widely used to detect cancer and to monitor cancer progression. They can be grouped into markers that are identified in cancer cells and markers that are secreted into body fluids. To perform early stage cancer diagnosis, the second group of markers is more appropriate. A promising method that allows minimal invasive tumor diagnosis based on markers is mass spectroscopy. Matrix-Assisted Laser Desorption and Ionization (MALDI) mass spectroscopy evaluated by 'peak probability contrasts' revealed an accuracy of around 70% for ovarian cancer (1). Similar approaches for pancreatic cancer performed slightly better with 88% sensitivity and 75% specificity (2).

Tumor antigens in blood sera represent an alternative approach for minimally invasive cancer diagnosis. A popular example is the prostate specific antigen (PSA) that is widely used in the diagnosis of prostate cancer (3). Since PSA is also present in the blood sera of 33% of unaffected people, PSA as a single tumor marker shows a lack of specificity. Likewise, other single antigen markers including CA-19.9 (pancreatic cancer) and CA-15.3 (breast cancer) show severe limitations (4). Recent studies strongly indicate that antigen marker sets significantly improve the specificity and sensitivity of cancer diagnosis compared to single antigen markers (5–7). Our Minimally Invasive Multiple Marker (MIMM) approach for meningioma (8) e.g. is based on 57 meningioma-associated antigens. Meningiomas are frequently occurring, generally benign intracranial tumors that are grouped by the World Health Organization (WHO) in three grades, grade I (common type), grade II (atypical) and grade III (anaplastic) meningioma. On a data set of 183 seroreactivity profiles from 83 meningioma and 90 normal sera, MIMM reached a specificity of 96.2% [95% confidence interval (CI) = (96.0–96.5%)], sensitivity of 84.5% (95% CI = 84.3–84.8%), and accuracy of 90.3% (95% CI = 90.1–90.4%). The area under the receiver operator curve (AUC-value) was 0.957 (95% CI = 0.956–0.957%).

We developed a web-based application, called 'Seroreactivity Profile Classification Service' (SePaCS) that gives experimental groups easy access to several supervised statistical learning approaches for classifying seroreactivity profiles. The results of SePaCS are summarized in an easy interpretable table that contains for each seroreactivity profile and each classification method the predicted class label. Our tool also provides a detailed

*To whom correspondence should be addressed. Email: ack@bioinf.uni-sb.de

result file containing for example, graphical representation of computed results. We demonstrate the capabilities and the ease-of-use of our web-based application on the example of meningioma.

MATERIALS AND METHODS

Supervised learning methods

We tested a variety of supervised learning methods on a meningioma data set. The approaches that yielded the best results were 'Naïve Bayes (NB) Classifiers' (NB), 'Support Vector Machines' (SVM) with radial basis kernel functions (9), 'Linear Discriminant Analysis' (LDA), and 'Diagonal Discriminant Analysis' (DLDA). Since further evaluations on glioma autoantibody profiles confirmed the results, these statistical learning approaches were included in SePaCS.

If the WHO grades of cancer sera in the data sets are also provided by the user, SePaCS additionally offers the possibility to predict the WHO grade of these sera using a modified 'NB Classifier' (NBC). All statistical computations are performed using the R language (10).

Mutual information MI

The mutual information (MI) is a well-known measure in information theory introduced by Shannon (11). The MI of an antigen *s* and the disease state (for example, cancer and control) represents a measure of the information content that *s* provides for the classification task, i.e. the disease state. Given two random variables X and Y, the MI I(X,Y) is a measure of the reduction in uncertainty about X due to the knowledge of Y. In our case, X and Y are binary random variables. The two possible states of the random variable X are 'normal' ($X=0$) or 'diseased' ($X=1$). In our application, the binary random variable Y represents the occurence of the antigen s, i.e. Y can take the states 's not detected' ($Y=0$) or 's detected' ($Y=1$). Thus, we can consider I(X,Y) as the reduction in uncertainty about the disease state due to the occurrence of antigen *s*. The higher the value of the MI of antigen s, the more 'valuable' *s* is for the classification task.

Subset selection

Variable selection is a widely used machine learning technique to increase the performance of classification methods by focusing on a subset of relevant features. Basically, two methods for feature subset selection exist, filter and wrapper approaches (12). Filter approaches perform the subset selection as a pre-processing step independent of the classification algorithm (13–15). One disadvantage of filter approaches is that two correlated features may be both included in the selected subset. In contrast, wrapper approaches conduct the search for an appropriate feature subset using the classification algorithm as part of the function for evaluating variable subsets (16), avoiding the problem of correlated features. Thus, in general wrappers should be preferred over filter approaches although they are computationally much more demanding.

We compared different wrapper and filter approaches that revealed comparable effectiveness. The subset selection method that showed the best performance is based on the so called 'MI' (11). By computing the MI of an antigen and the class label (0 for normal and 1 for cancer sera), it allows for measuring the diagnostic information that the antigen provides. We use a greedy algorithm that adds in each step the antigen that provides the highest MI. Using 10-fold cross validation we determine the subset that shows the lowest error rate for classification.

SEROREACTIVITY PROFILES CLASSIFICATION SERVICE

SePaCS offers two different modes of operation. In the first mode, usage of own training data, the user can upload two antibody profile sets, a training and a test set. In the second mode, no training data set has to be provided. Instead, classification methods that are already trained on our data sets can be applied to the uploaded antibody profiles. Currently, SePaCS provides trained classifiers for meningioma. Similar models for other cancer entities will be available soon, starting with predictors for gliomas (manuscript in preparation). Additionally, we plan to provide classification methods trained with autoantibody profiles of prostate cancer patients, nephroblastoma patients, and patients with lung cancer.

In both operating modes the results of SePaCS are summarized in tabular form, i.e. for each classification method and each test sample the result table contains a '1' if a tumor is predicted and a '0' otherwise.

The web-interface of SePaCS is implemented in Perl and consists of three modules: parameter specification, data upload, and data processing and output. The required parameters can be specified using the web-interface. On this interface, the user has to select the operating mode and has to choose at least one of the offered classification methods. At present, the user can choose from the following statistical learning methods: four different versions of a NB Approach, SVM with a radial basis function kernel (9), LDA and DLDA. Furthermore, the user has to define a parameter, specifying, whether a subset selection based on MI (11) should be performed

In the second step the data is uploaded. If the user intends to train the selected statistical learning algorithms with his own training set (operating mode one), he can optionally assign names to the antigens. The antigen names have to be separated by a semicolon. If no names are given, the antigens are numbered. Afterwards, the user has to upload the training data that should be imported as a matrix M of size $n \times (p+1)$, where each of the n rows represents one serum. The first column denotes the class label, i.e. a '0' for each normal serum and a '1' for each patient's serum. Each of the following p columns represents an antigen. The matrix entry $M[i,j+1]$ contains the information whether antigen j has been detected in serum i ($M[i,j+1]=1$) or not ($M[i,j+1]=0$). The entries of the data matrix are delimited by white spaces. The described format allows for an easy data-upload by 'copy and paste' from spreadsheets. An example for a training

data matrix M is provided in the supplemental material. In both operating modes, the antigen profiles to be classified are uploaded next. If m sera are to be diagnosed, this data matrix is expected to be of dimension $m \times p$, i.e. the matrix has one row per serum and one column per antigen. These sera can also be named. If sera names are given, they have to be separated by a semicolon, if no names are given, the test sera are numbered.

Additionally, SePaCS offers the option to upload a data file if a user intends to use own seroreactivity profiles. For details on the data file, we refer to the SePaCS tutorial, where an example data file can be downloaded.

The statistical analysis starts with an evaluation of the training data, including among others the mean antigen reactivity in cancer and control sera, the MI of single antigens, and the estimation of the classification methods' performance using standard 10-fold cross validation. Considering the MI profile, users can easily detect the most 'valuable' antigens that are especially suited to perform an accurate classification. The cross-validation error rates enables researchers to assess the classification results obtained for the test set. If a data set shows a low cross-validation error, the predictions of the test data are likely to be correct. In contrast, if a high cross-validation error rate is reached (maybe due to noisy data), the classification results of the test data may be incorrect. Thus, the first part of the statistical analysis facilitates the interpretation of the data set.

Thereafter, the supervised statistical learning methods are applied to the antibody profiles and the classification results are provided on a web-page starting with a summarizing table. This table shows the output for each classification method and each antibody profile. Positive predictions (tumor) are colored red and negative (normal) predictions are colored green. An example of a table is shown in Figure 1. The web-page additionally contains links to supporting plots, e.g. MI profiles.

Methode	1	2	3	4	5	6	7	8	9	10
NBA	0	0	0	0	0	1	1	1	1	1
NBA MI	0	0	0	0	0	1	1	1	1	1
NBB	0	0	0	0	0	1	1	1	1	1
NBB MI	0	0	0	0	0	1	1	1	1	1
NBD	0	0	0	0	0	1	1	1	1	1
NBD MI	0	0	0	0	0	1	1	1	1	1
LDA	0	0	0	0	0	1	1	1	1	0
LDA MI	0	0	0	0	0	1	1	1	1	1
DLDA	0	0	0	0	0	1	1	1	1	1
DLDA MI	0	0	0	0	0	1	1	1	1	1
SVM	0	0	0	0	0	1	1	1	1	1
SVM MI	0	0	0	0	0	1	1	1	1	1

Figure 1. Classification results of all supported supervised learning methods with and without subset selection for 10 randomly selected sera (5 control sera and 5 meningioma sera) based on the models trained with the meningioma data set. All methods showed correct classification results, only LDA without subset selection miss-classified a single meningioma serum as control serum.

Besides this summary page, details of the analysis are provided as PDF report. This report can be either downloaded or accessed online via a unique job ID. For example, the report generated with the meningioma data set described in 'Results' can be accessed by using the job ID 000001. The PDF report is divided into up to five sections, depending on the chosen parameters. In the first section, a summary of the analyzed data set is presented, including images of the data matrices as well as basic statistics of the training data set. For example, the mean antigen reactivity of healthy and diseased sera and a balloon plot of the antigen distribution are shown. An example of such a balloon plot for the meningioma data set is given in Figure 2. The second section contains the classification results for each serum and each classification method. For some of the statistical learning methods, as the NB approaches, additional graphical output is provided. Here, the quotient of the probabilities that a serum is a normal serum and that the serum is a cancer serum is plotted. An example of such a plot is shown in Figure 3. Test and training set are divided by the vertical blue line, and all sera above the horizontal green line are classified as cancer sera. If a subset selection based on MI has been performed, the next section presents the MI of all antigens. This section also contains the performance of the classification methods that have been evaluated on the training data using 10-fold cross validation as function of the subset size. Additionally, the classification results computed with the shrunken subsets are provided. In the last section, all available classification results are summarized in tabular form.

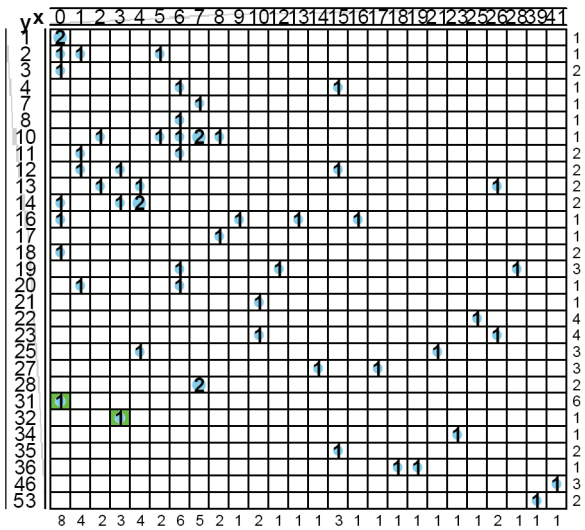

Figure 2. Analysis of the antigen distribution in the meningioma test set. The balloon plot shows at position i, j how many antigens occur in i meningioma sera and j control sera. The size of the blue balloons is chosen proportional to the number of antigens. The numbers on the right side of the figure and below the figure denote the sums of rows and columns. Antigens with high information content are typically located in the lower left corner of the balloon plot. The green highlighted antigens provide the highest diagnostic value (see also Figure 4). The antigens in the first column are only detected in meningioma sera and are called specific sera.

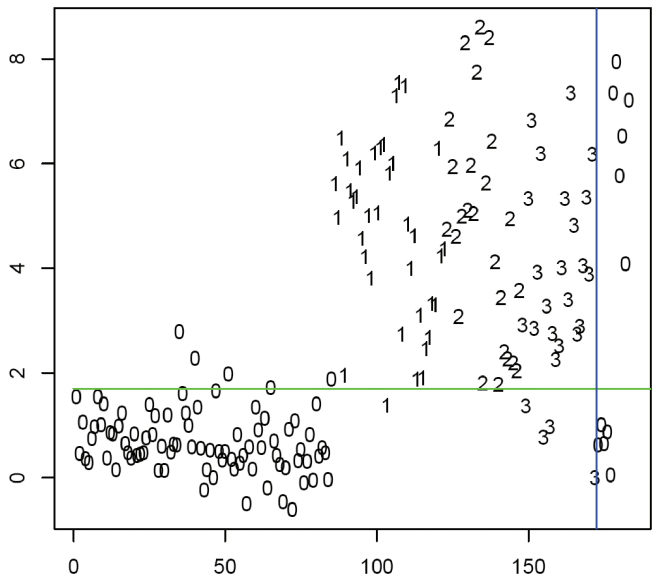

Figure 3. Classification result of the first NB approach. Ten sera are classified based on the model generated with 173 meningioma and control sera. Training and test sera are separated by a vertical blue line. All sera above the horizontal green line are predicted to be meningioma sera. Each zero in the training set represents a control serum, each number represents a serum of a meningioma serum with the respective WHO grade. The *y*-axis represents the logarithmized quotient P(A) of P(M|A) divided by P(N|A), where P(M|A) is the conditional probability that given the antibody profile A the serum is a meningioma serum, and P(N|A) is the conditional probability that given the antibody profile A the serum is a normal serum. The higher the value P(A) the more probable the serum is a meningioma serum.

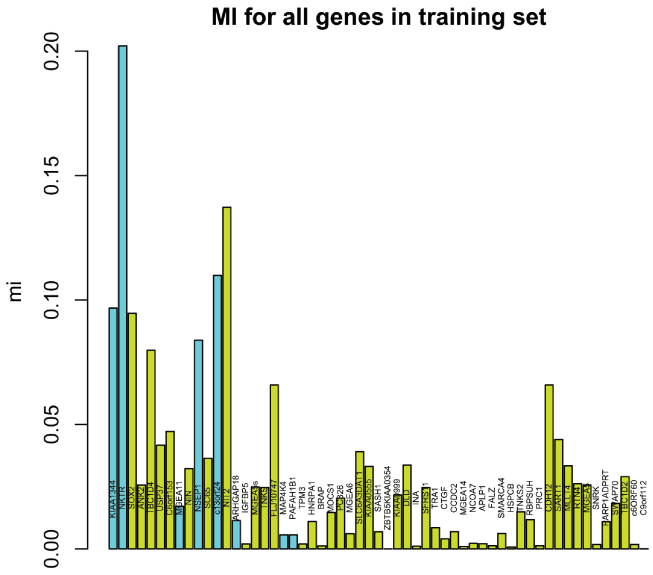

Figure 4. MI of the 57 antigens and the disease state. The color of bars indicates whether the antigens occur only in diseased sera (so called specific antigens; blue bars) or in diseased and control sera (green bars). That representation allows users to rank the antigens easily according to their diagnostic value. In general, specific antigens show higher information content than non-specific antigens, however, some non-specific antigens as NIT2 also provide a high diagnostic value.

RESULTS

The functionality of SePaCS is demonstrated exemplarily on a meningioma data set including a total of 183 sera (90 meningioma patients and 93 controls). These sera have been screened for reactivity against 57 antigens that are known to be meningioma associated (6). For example, we divided the data set such that classification was performed for 10 randomly selected sera based on a training data set of 173 sera. The pdf report for this example can be reviewed on www.bioinf.uni-sb.de/sepacs, using the job ID 000001 and is also available in the supplementary material.

The analysis of the training set showed that the mean antigen reactivity in meningioma sera (20%) is significantly higher than in control sera (11%). The antigen distribution shown in Figure 2 reveals that in meningioma sera up to 53 of the 57 considered antigens have been detected, whereas in normal sera only up to 41 antigens have been found. Out of the 57 antigens 8 antigens react only with meningioma sera, but not with control sera as detailed in Figure 2. In Figure 4, the MI of all antigens is provided. The diagnostic value of each antigen can be directly compared to the value of all other antigens. The antigen with the highest MI (0.2) is NKTR that occurs in 31 of 87 meningioma sera (37%), but not in a single control serum. NKTR represents the last antigen in the first column of the balloon plot in Figure 2. The antigen with the second highest MI (0.14), NIT2, is detected in 3 of 85 control sera and in 32 of 87 meningioma sera. The two antigens providing the highest MI are highlighted in Figure 2.

Table 1. Cross validation error rates of the different classification methods together with the antigen subset size

Method	Subset size	Error rate
NBA	31	0.087
NBB	40	0.081
NBC	32	0.283
NBD	34	0.081
Linear Discriminant Analysis	32	0.104
Diagonal Discriminant Analysis	40	0.104
Support Vector Machine	50	0.087

Table 1 shows the cross validation error rates of the training set for the subset selection together with the respective subset sizes. This information enables the user to judge the predictive power of the different classification methods regarding his data set. In our example, best performance is obtained with the NB Classifiers, whereas LDA and DLDA show significantly decreased performance.

As shown in Figure 1, the first five sera that stem from control persons are classified by all prediction methods as non-meningioma sera. With the exception of LDA (without subset selection) that misclassifies one meningioma serum, all meningioma sera have been correctly predicted. We also tested the server with seroreactivity patterns of 95 glioma sera versus 82 control sera. The results were of comparable quality as the results computed by using meningioma antibody profiles.

DISCUSSION

SePaCS grants non-experts in the field of statistical learning easy access to a comprehensive analysis framework for classifying seroreactivity profiles. Our tool offers the possibility to analyze seroreactivity profiles not only from different tumor entities, but from a wide variety of other human diseases that trigger a complex immune response, e.g. autoimmune diseases. Although, SePaCS was designed to analyze autoantibody profiles, it can be used for any kind of binary data.

The easy usage of our statistical framework and its diagnostic value have been demonstrated with a meningioma data set. A second test with glioma seroreactivity profiles showed a similar performance. Currently, we are preparing analyzes of seroreactivity profiles for other tumor types.

ACKNOWLEDGEMENTS

Funding to pay the Open Access publication charges for this article was provided by internal funds of the Saarland University.

Conflict of interest statement. None declared.

REFERENCES

1. Tibshirani,R., Hastie,T., Narasimhan,B., Soltys,S., Shi,G., Koong,A. and Le,Q. (2004) Sample classification from protein mass spectrometry, by'peak probability contrasts'. *Bioinformatics*, **20**, 3034–3044.
2. Koomen,J., Shih,L., Coombes,K., Li,D., Xiao,L., Fidler,I., Abbruzzese,J. and Kobayashi,R. (2005) Plasma protein profiling for diagnosis of pancreatic cancer reveals the presence of host response proteins. *Clin. Cancer. Res.*, **11**, 1110–1118.
3. Vicini,F.A., Vargas,C., Abner,A., Kestin,L., Horwitz,E. and Martinez,A. (2005) Limitations in the use of serum prostate specific antigen levels to monitor patients after treatment for prostate cancer. *J. Urol.*, **173**, 1456–1462.
4. Sidransky,D. (2002) Emerging molecular markers of cancer. *Nat. Rev. Cancer.*, **2**, 210–179.
5. Wang,X., Yu,J., Sreekumar,A., Varambally,S., Shen,R., Giacherio,D., Mehra,R., Montie,J.E., Pienta,K.J. *et al.* (2005) Autoantibody signatures in prostate cancer. *N. Engl. J. Med.*, **353**, 1224–1235.
6. Comtesse,N., Zippel,A., Walle,S., Monz,D., Backes,C., Fischer,U., Mayer,J., Ludwig,N., Hildebrandt,A. *et al.* (2005) Complex humoral immune response against a benign tumor: Frequent antibody response against specific antigens as diagnostic targets. *Proc. Natl. Acad. Sci. U S A*, **102**, 9601–9606.
7. Erkanli,A., Taylor,D., Dean,D., Eksir,F., Egger,D., Gayer,J., Nelson,B., Stone,B., Fritsche,H. *et al.* (2006) Application of Bayesian modeling of autologous antibody responses against ovarian tumor-associated antigens to cancer detection. *Cancer res.*, **66**, 1792–1798.
8. Keller,A., Comtesse,N., Ludwig,N., Hildebrandt,A., Meese,E. and Lenhof,H.P. (2006) a minimally invasive multiple marker approach allows highly efficient detection of low-grade tumors. *BMC Bioinformatics*, **7**, 539.
9. Vapnik,V. (1995) *The nature of statistical learning theory* Springer.
10. R Development Core Team. (2006) R: a language and environment for statistical computing
11. Shannon,C.E. (1948) A Mathematical theory of communication. *The Bell System Technical J.*, **27**, 623–656.
12. John,G.H., Kohavi,R. and Pfleger,K. (1994) Irrelevant features and the subset selection problem. In *Proceedings of the International Conference on Machile Learning*. pp. 121–129.
13. Allmuallim,H. and Deitterich,T.G. (1991) Learning with many irrelevant features. *Proceedings of the Ninth National Conference on AI*. 547–552.
14. Holmes,G. and Nevill-Manning,C.G. (1995) Feature selection via the discovery of simple classification rules, *Proceedings International Symposium on Intelligent DataAnalysis* (IDA-95).
15. Kira,K. and Rendell,L. (1997) A practical approach to feature selection, *Procedings of the Ninth International Conferenceon ML*. 249–256.
16. Kohavi,R. and John,G.H. (1997) Wrappers for Feature Subset Selection. *Artificial Intelligence*, **97**, 273–324.