

External Cross-Validation for Unbiased Evaluation of Protein Family Detectors: Application to Allergens

Daniel Soeria-Atmadja,¹ Mikael Wallman,² Åsa K. Björklund,¹ Anders Isaksson,² Ulf Hammerling,¹ and Mats G. Gustafsson^{2,3*}

¹Division of Toxicology, National Food Administration, Uppsala, Sweden

²Department of Genetics and Pathology, Uppsala University, Rudbeck Laboratory, Uppsala, Sweden

³Department of Engineering Sciences, Uppsala University, Uppsala, Sweden

ABSTRACT Key issues in protein science and computational biology are design and evaluation of algorithms aimed at detection of proteins that belong to a specific family, as defined by structural, evolutionary, or functional criteria. In this context, several validation techniques are often used to compare different parameter settings of the detector, and to subsequently select the setting that yields the smallest error rate estimate. A frequently overlooked problem associated with this approach is that this smallest error rate estimate may have a large optimistic bias. Based on computer simulations, we show that a detector's error rate estimate can be overly optimistic and propose a method to obtain unbiased performance estimates of a detector design procedure. The method is founded on an external 10-fold cross-validation (CV) loop that embeds an internal validation procedure used for parameter selection in detector design. The designed detector generated in each of the 10 iterations are evaluated on held-out examples exclusively available in the external CV iterations. Notably, the average of these 10 performance estimates is not associated with a final detector, but rather with the average performance of the design procedure used. We apply the external CV loop to the particular problem of detecting potentially allergenic proteins, using a previously reported design procedure. Unbiased performance estimates of the allergen detector design procedure are presented together with information about which algorithms and parameter settings that are most frequently selected. *Proteins* 2005;61:918–925. © 2005 Wiley-Liss, Inc.

Key words: protein classification; bias; bioinformatics; supervised learning; allergy; sequence alignment

INTRODUCTION

In protein science and computational biology, key issues are design and evaluation of detectors for proteins of a specific family. Such families can be defined on structural, evolutionary, or functional grounds and many algorithms aimed at the specific detection of certain protein groups, each with similar characteristics, are already available. Examples thereof include the detection of:

1. complete amino acid sequences of transmembrane proteins or fractions thereof,^{1–3}
2. protein targeting signals,^{4–7}
3. subcellular location,^{8–14}
4. RNA-binding proteins,¹⁵
5. protein structure,¹⁶
6. alpha-turns,¹⁷
7. subgroups of G-protein coupled receptors,¹⁸
8. homodimers,¹⁹
9. subgroups of enzymes,²⁰ and
10. allergens.^{21–31}

In addition to these topics, there are also studies on less application-specific approaches.^{32–34}

As can be understood from the many applications listed above, reliable detection of protein families is an invaluable help in many areas of biology and biomedicine. In essentially all of these applications, it is important to obtain unbiased estimates of the detection performance so that overly optimistic interpretations of results are avoided. Potential sources of bias are, however, easily overlooked, particularly because the datasets available for analysis usually are rather small. Among the broad range of reports on protein family detection, it is often difficult to determine to what extent the various sources of bias are avoided. This applies to performance estimates obtained by any validation technique, such as holdout test, *n*-fold cross-validation (CV), or different bootstrap methods. In

Abbreviations: CV, cross-validation; LBDet, lower boundary for detection; UBFA, upper boundary for false alarms.

The Supplementary Material referred to in this article can be found at <http://www.interscience.wiley.com/jpages/0887-3585/suppmat/>

Grant sponsor: The Cancer and Allergy Fund (Stockholm); Grant sponsor: Carl Tryggers Stiftelse (Stockholm); Grant sponsor: the Göran Gustafsson foundation (Stockholm); Grant sponsor: the Faculty of Science and Technology (Uppsala University); Grant sponsor: the Beijer foundation (Uppsala); Grant sponsor: the Wallenberg Consortium North; Grant sponsor: the Magnus Bergvall foundation; Grant sponsor: the Marcus Borgström foundation; Grant sponsor: the Swedish Society for Medical Research

Åsa K. Björklund's present address is Stockholm Bioinformatics Center, SCFAB, Stockholm University, SE-106 91 Stockholm, Sweden.

*Correspondence to: Mats Gustafsson, Department of Engineering Sciences, Uppsala University, P.O. Box 528, SE-751 20 Uppsala, Sweden. E-mail: mg@signal.uu.se

Received 1 March 2005; Accepted 17 May 2005

Published online 17 October 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20656

n -fold CV, all data examples are partitioned in n roughly equally sized fractions. In each of the n iterations, each fraction is used for validation purposes, whereas the remaining $n-1$ fractions are used to design the detector. The final estimate is calculated as the average estimate of the n iterations. If n equals the total amount of data N , the procedure is often referred to as leave-one-out (LOO) CV.

It is well known in supervised learning that performance evaluation of a detector, using test examples that already have been involved in *any* part of the design, may result in appreciably over-optimistic performance estimates.³⁵ The most apparent course is to use all examples for both design and test. Although this overtly inappropriate evaluation is rarely seen in the literature, there are less obvious cases, e.g., allowing all data (instead of exclusively design data) to be used in the feature selection step, before detector design and error rate estimation. The potential bias caused by using test examples in the feature selection procedure seems to have been overlooked in many areas but was recently pointed out in a study on classification of microarray gene expression data.³⁶ The study showed that negligence of this problem might result in quite dramatic biases, in particular for small sample sizes.

In this article, we deal with another often overlooked bias in the context of protein family detection, which is associated with parameter optimization based on the error estimates. Validation techniques are correctly employed when they are used for unbiased performance estimation of a single parameter setting, or used to select the most promising parameter setting (being that of the lowest error estimate) from several alternatives. Even though the parameter setting obtained in the latter case has been appropriately selected, the entailed error estimate may, however, show an overly optimistic error rate, as discussed and demonstrated in this work. Parameter tuning, either for optimizing parameter settings in the feature selection process or in the detection process, increases the risk of selecting a parameter setting that is overfitted to data presented in the validation procedure. In other words, a bias can easily appear because the selected combination of detector algorithm and parameter setting is not being evaluated on any independent test examples outside the validation procedure used for parameter selection. An illustrative practical example of such a bias was recently reported by Bhasin and Raghava³⁷ in the context of detecting HLA-DRB1*0401-binding peptides in antigen sequences. The detector performance drastically decreased, when evaluated on a completely independent external holdout test set, as compared with the CV estimate selection of detector algorithm and parameter settings.

In this work, we show that it is possible to obtain unbiased performance estimates, while allowing for detector setting optimization, by means of introducing an external CV loop that embeds an internal validation loop. The internal validation is strictly used for parameter optimization and the accordingly selected parameter setting is subsequently evaluated on test data in the external

CV loop, which are independent of data used in the internal validation procedure. In the internal validation, we use CV for comparing detector settings but it could also be replaced, for example, by a bootstrap or holdout test procedure. A similar approach has been used to validate multivariate regression models of structure-activity relationships in chemistry and to validate the potential of proteochemometrics.^{38,39} By means of computer simulations, we illustrate the power of the external CV procedure and we thereafter exemplify how it can be used practically by applying it to a previously reported detector of potential protein allergens.

MATERIALS AND METHODS

External CV Loop for Identification and Avoidance of Validation Bias on Simulated Data

Computer simulations were performed to illustrate that the performance estimate, associated with the best parameter found among several alternatives, is biased. One thousand different simulated datasets, each consisting of 25 examples per class, were generated from two partially overlapping Gaussian distributions. For each of the generated datasets, the 25 examples per class were partitioned into 10 subsets consisting of two or three samples each. In each iteration of an external 10-fold CV procedure, nine subsets (45 examples) were used to form an external design set and the remaining subset (five examples) was kept as a small external test set. In each of the 10 iterations, an internal LOO error estimate was determined for all possible choices of the number of neighbors (k) of the supervised k -nearest neighbor (kNN) classifier. For more details of the kNN, see Hastie et al.³⁵ The LOO error rate estimate was *strictly* based on the 45 data samples in the external design set, i.e., the five examples in the external test set was not included in this estimation. The value of k , which yielded the lowest LOO error rate estimate, was finally used to design a kNN classifier using all the examples in the (external) design set. In the event of a draw between two or more values of k , the least complex classifier was chosen, i.e., the one with the smallest value of k . The resulting classifier was then tested on the external test set as well as on a large independent validation set consisting of 10,000 samples (5,000 per class). Because the procedure was iterated for all external design- and validation sets, 10 selected classifiers were obtained. Thus, for each of these 10 selected classifiers, we obtained one error rate estimated from the internal LOO procedure, one estimate from a small external test set, and one estimate from a large external test set (5,000 examples per class). Under the assumption that 5,000 samples from each class are sufficient, the obtained 10 estimates from the large test set are measures of the classifiers' "true" error rates. The three error rate estimates will further on be referred to as "internal," "external," and "true," respectively. For each of the three error rates, a mean value was calculated based on the 10 selected classifiers. This external CV procedure was then repeated 1,000 times, each time using one of the generated datasets.

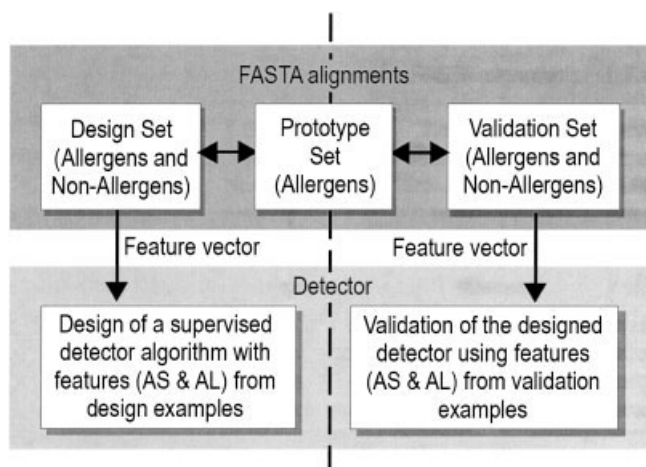


Fig. 1. Outlines of the detector of potential protein allergenicity. The figure shows the general outlines for the design and validation of the allergen detector. Design sequences (allergens and non-allergens) are aligned against the prototype sequences, using FASTA3. Alignment score (AS) and alignment length (AL) from these alignments are compiled to a feature vector, which is used to design a supervised detector algorithm. The accordingly designed detector is subsequently validated with a similar feature vector corresponding to the validation sequences.

Design and Validation of Allergen Detectors

The assessment scheme for evaluation of different allergen detectors (detector settings) in the work reported here largely adheres to our earlier work.²⁹ First, the allergen and non-allergen datasets were organized into subsets. The allergen dataset was randomly split into a validation set and a master set. Subsequent to alignment-based clustering of the master set sequences, one example from each such cluster was randomly selected into a prototype set, whereas the remaining part constituted the design set. This process ensured that the prototype set was as diverse as possible in the context of sequence similarity. The non-allergen dataset was randomly divided into a validation set and a design set. The steps for detector design and validation are illustrated in Figure 1 and can be summarized as follows: Local alignment procedures were performed between design sequences (allergen and non-allergen) and prototype sequences. The resulting alignment scores (ASs) and the alignment lengths (ALs) were subsequently used as input features to design a detector of allergenic potential, using a supervised detector algorithm. The accordingly designed detector was finally validated with ASs and ALs from alignments between validation sequences against the prototype sequences (allergens and non-allergens).

The allergen detector settings were selected from the following alternatives: 1) the number of prototypes (100, 150, or 200), 2) the substitution matrix (Blosum80 or Blosum50), 3) the gap penalties gap opening/gap extension (default; $-12/-2$ or high; $-10^6/-10^6$), 4) the supervised detector algorithm (kNN, LG, or QG). For details on these supervised algorithms, see Hastie et al.³⁵ For kNN, two different values of the number of neighbors, k (7 and 9), were tested. Thus, $3 \times 2 \times 2 \times 4 = 48$ detector settings were evaluated.

Two hundred eighty-three and 871 redundant allergen and non-allergen amino acid sequences (or sequence clusters), respectively, were used in the study. The sequences were selected from earlier reported datasets consisting of 578 allergens³¹ and 1,007 non-allergens,²⁹ which both are publicly available at the following Web site: <http://www.slv.se/allergyalign/tables2>. More information on the datasets and the procedure used to avoid redundancy is available in the Supplementary Material.

Application of the External CV Procedure on Detection of Potential Allergenicity

The external CV loop was applied to the aforementioned detector of potentially allergenic proteins and the procedural steps in each external iteration i are illustrated in Figure 2. A tenth of all sequences was held out as the external Validation set i (Allergen and Non-Allergen) whereas the remaining examples constituted the external Allergen Master set i and the external Non-Allergen Design set i , respectively. These two sets were first used in an internal 10-fold CV to find the most promising detector setting from 48 alternatives (each setting corresponds to a unique combination of algorithm and parameter values) and were subsequently used externally to design a final detector i . For more information on the organization of datasets in both the external and internal CV iterations, see Figure S1 in the Supplementary Material.

For each of the 10 external CV iterations i , an internal 10-fold CV was used for selection of detector settings. The selection procedure was performed using *conservative* receiver operating characteristics (ROC) curves based on the 10 internal CV iterations ij . An ordinary ROC curve depicts the probability of detection (sensitivity, correctly classified allergens) on the ordinate versus the probability of false alarm (1-specificity, erroneously classified non-allergens) on the abscissa. Each point on the curve shows the variation of the two aforementioned measures with changes in the decision threshold, i.e., the trade-off between sensitivity and specificity for a two-class test.⁴⁰ In contrast to the *point* estimates of the probability of detection and false alarm as shown in an ordinary ROC curve, a conservative ROC curve shows the *lower* (95%) confidence boundary for detection (LBDet) versus the *upper* (95%) confidence boundary for false alarms (UBFA).³¹ For each of the 48 detector settings, the confidence boundaries LBDet and UBFA were calculated from the 10 internal estimates obtained in the internal CV loop. The setting yielding the highest LBDet for the majority of the UBFA levels, ranging from 1 to 25%, were selected for external evaluation. This range encompasses an appreciable part of the conservative ROC curves and false alarm levels higher than 25% are beyond the scope of practical importance. This majority voting in the selection was used to obtain a robust setting in comparison to a setting based on the maximal LBDet at a single UBFA level.

In each iteration i of the external 10-fold CV procedure, a final design and evaluation was ultimately performed. In conjunction with the accordingly selected detector setting, the two datasets external Allergen Master set i and

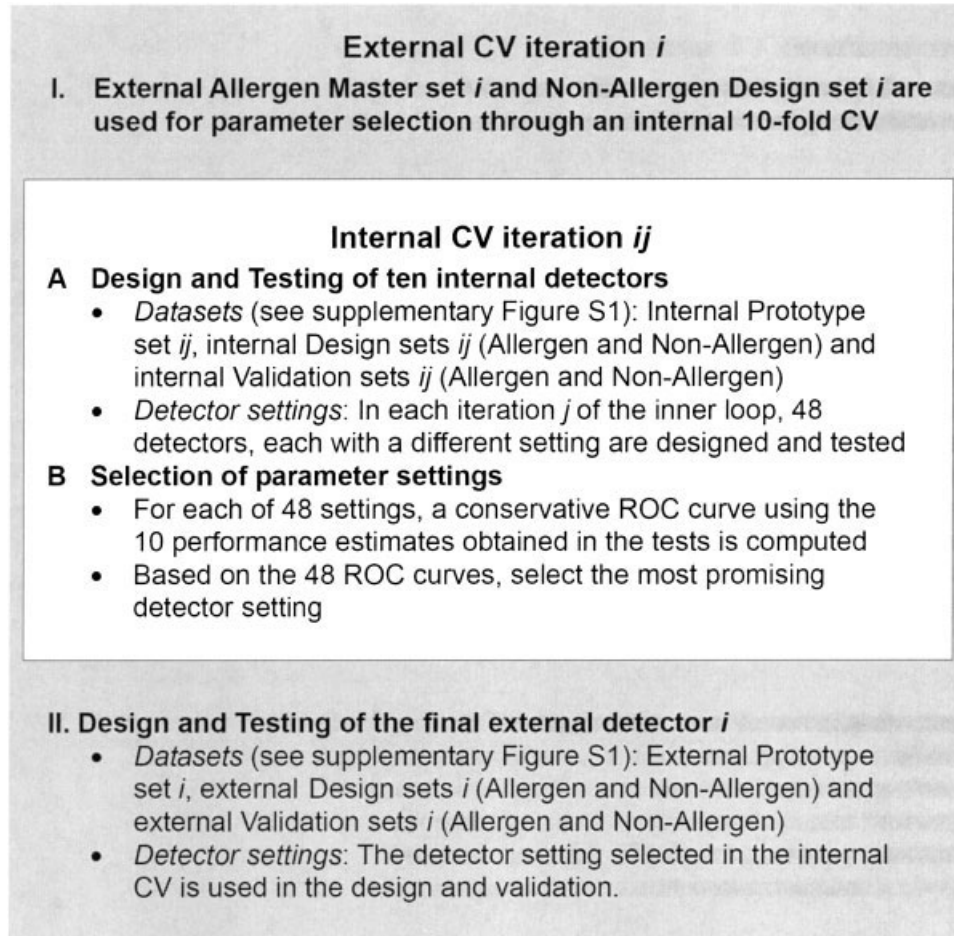


Fig. 2. External CV procedure applied to the design and evaluation of an allergen detector. Internal CV iteration ij (in white) for detector selection and external CV iteration i (in gray) for detector evaluation. The figure depicts the procedure of using 10 internal CV iterations j for each of 10 external CV iterations i to separate the evaluation of detector performance estimates from the selection of detector setting. In the external evaluation of detector i , the detector is designed with Prototype set i and Design sets i (allergen and non-allergen) using the detector setting suggested most suitable according to the 10 internal CV:s. The performance of the ultimately designed detector i is validated with the held-out external Validation sets i (allergen and non-allergen).

external Non-Allergen Design set i were used to design an allergen detector i , whose performance was evaluated with the sequences in the external Validation set i .

Computer Programming

Except for the local and global alignment algorithms included in the FASTA3 package,⁴¹ development and computational exercises were performed in MATLAB (The MathWorks, Inc., Natick, MA). Particular extension modules used were the Statistical Toolbox from MathWorks and the PRTTools toolbox.⁴²

RESULTS

Computer Simulations for Identification of CV Bias

Computer simulations were performed as described in M&M, using data generated from two Gaussian distributions. For each of the 1,000 simulated datasets and for each type of error estimate (internal, external, true), 10 corresponding error rate estimates were obtained. The averages of these 10 estimates were computed, thus result-

ing in 1,000 different sets of mean values for each error, denoted m_{Int} , m_{Ext} , and m_{True} . For each simulation, m_{Int} , m_{Ext} , and m_{True} were combined into joint observations consisting of three possible pairs: $(m_{\text{Int}}, m_{\text{True}})$, $(m_{\text{Ext}}, m_{\text{True}})$, and $(m_{\text{Int}}, m_{\text{Ext}})$, which are presented as two-dimensional histograms in Figure 3. The averages of the 1,000 m_{Int} , m_{Ext} , and m_{True} , denoted M_{Int} , M_{Ext} , and M_{True} , are listed in Table I. Also listed in Table I are the standard deviations of the means M_{Int} , M_{Ext} , and M_{True} . As expected and shown from Figure 3(A, B), as well as in Table I, the internal average error estimate M_{Int} has a large optimistic bias whereas the external average error estimate M_{Ext} is almost identical to the true average error estimate M_{True} . Moreover, results in Figure 3(C) show that the internal and external estimates have a strong positive correlation, but the large variances imply that the result from a single study (representing one of these 1,000 points) might be quite misleading, even if it is unbiased. For example, as can be seen in Figure 3(B), the vast majority of the average of "true" error rate estimates are observed on

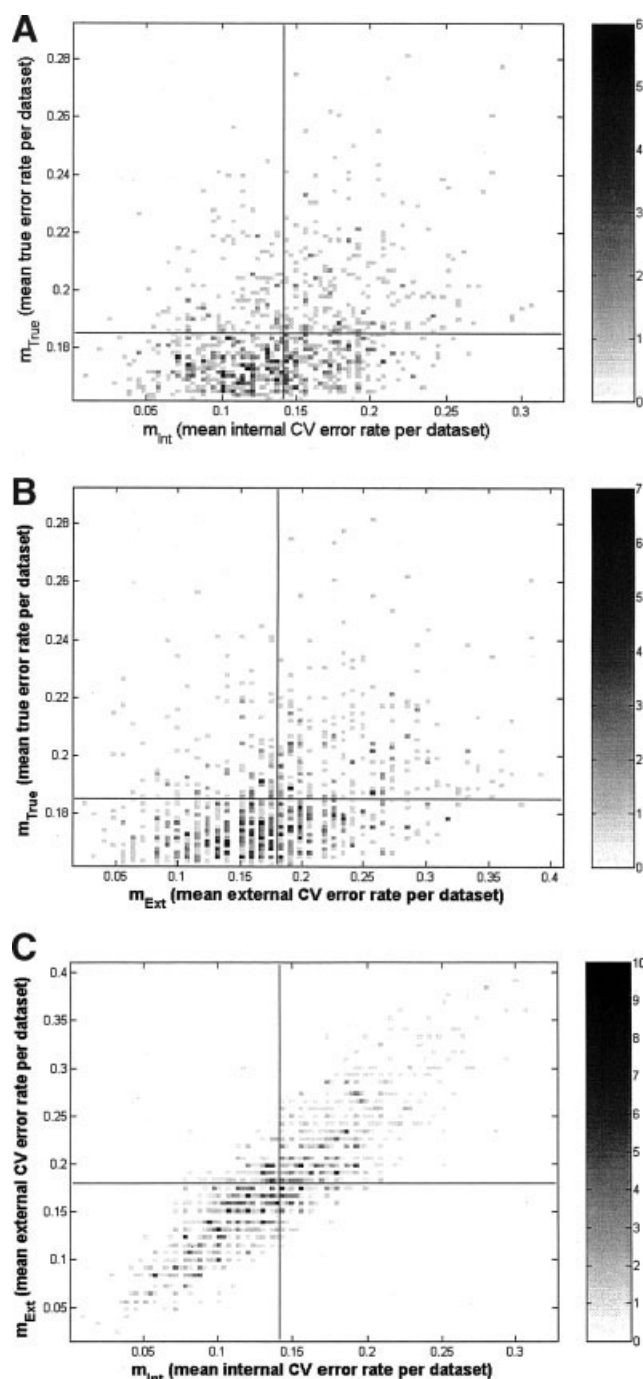


Fig. 3. Comparison of internal and external error rate estimates with true estimates. Results from computer simulations are illustrated as two-dimensional histograms over pairs of the 100 m_i ($i = Int, Ext, or True$), which are the averages of the 10 error rate estimates, obtained in the external 10-fold CV. The histograms were created by dividing the range between the largest and smallest value on each axis into 100 equally sized sections, thus forming a two-dimensional grid. For each square in the grid, the number of observations was counted. The resulting matrices were finally visualized by giving each square a gray-scale value corresponding to the number of points it contained, spanning from white (no points) to black (many points). The center of each distribution, which is the average of the 1,000 averages m_i , is denoted M_i and is indicated by a black line. **A:** The average true error rate m_{True} versus the average internal error rate m_{Int} . **B:** The average true error rate m_{True} versus the average external error rate m_{Ext} . **C:** m_{Ext} versus the average internal error rate m_{Int} .

TABLE I. Averages (M) and Standard Deviations of Averages (STD) Based on 1,000 Average Error Rate Estimates

Type of error rate i	M_i	STD_i
Int (internal)	0.141 ^a	0.0015 ^a
Ext (external)	0.180 ^b	0.0020 ^b
True	0.1863 ^c	0.0006 ^c

^aEach of the 1,000 internal error rate estimates is a LOO estimate.

^bEach of the 1,000 external error rate estimates is a 10-fold CV estimate (on five samples).

^cEach of the 1,000 true error rate estimates is an average of 10 holdout estimates (on 10,000 samples).

the interval [0.16, 0.22], whereas the corresponding “external” error rate estimates are observed on the interval [0.05, 0.30].

Selection of Detector Settings for Detection of Potential Allergens

Table II lists the selected detector settings that were selected as the most promising in the internal CV and used in the subsequent design of the 10 external allergen detectors i . Whenever several sets of detector settings appeared as being equivalent, one of them was selected at random. Regarding the selection of detector algorithms, results between each external CV were consistent, showing a preference for the kNN algorithms (either $k = 7$ or 9) over both the LG and the QG counterparts. Moreover, a rather clear preference was evident for a combination of Blosom80 with a high gap penalty setting. No clear-cut difference on prototype numbers was, however, seen; 100, 150, and 200 prototypes were chosen at least two times each.

Evaluation of Detector Performance

The selected detector settings were finally evaluated with the corresponding external Validation sets i (Allergen and Non-Allergen), using Allergen Master set i and Non-Allergen Design set i for design. Hence, the external Master set i and the Non-Allergen Design set i were in this process used strictly for design (see M&M, Fig. 2 and supplementary Fig. S1). The corresponding performance for each of these 10 external detectors i is depicted in Figure 4, using ordinary ROC curves based on point estimates of the error rates. In addition to the ROC curves, the average probability of detection (of the 10 detectors i) in each false alarm point is depicted. The average ROC curve in the allergen detector design procedure is analogous to one of the 1,000 average values m_{Ext} in the computer simulations.

DISCUSSION

Bias Attributed to Improper Validation for Classifier/Detector Performance Estimation

In a recent report on prediction of protein domains, Liu and Rost⁴³ performed a 10-fold CV procedure in which approximately 9% of the examples were set aside as an external holdout set and the remainder for design and

TABLE II. Top-Ranked Detector Settings to Be Used in Design of Each External Allergen Detector i

CV iteration	Setting number	Detector algorithm	Amount of prototypes	Substitution matrix	Gap penalties (opening/extension)
1	36	kNN ($k = 9$)	200	Blosum80	High ($-10^6/-10^6$)
2	40	kNN ($k = 7$)	100	Blosum80	High ($-10^6/-10^6$)
3 ^a	36	kNN ($k = 9$)	200	Blosum80	High ($-10^6/-10^6$)
3 ^a	48	kNN ($k = 7$)	200	Blosum80	High ($-10^6/-10^6$)
4	36	kNN ($k = 9$)	200	Blosum80	High ($-10^6/-10^6$)
5	32	kNN ($k = 9$)	150	Blosum80	High ($-10^6/-10^6$)
6	28	kNN ($k = 9$)	100	Blosum80	High ($-10^6/-10^6$)
7 ^a	28	kNN ($k = 9$)	100	Blosum80	High ($-10^6/-10^6$)
7 ^a	32	kNN ($k = 9$)	150	Blosum80	High ($-10^6/-10^6$)
7 ^a	36	kNN ($k = 9$)	200	Blosum80	High ($-10^6/-10^6$)
8	28	kNN ($k = 9$)	100	Blosum80	High ($-10^6/-10^6$)
9	48	kNN ($k = 7$)	200	Blosum80	High ($-10^6/-10^6$)
10	32	kNN ($k = 9$)	150	Blosum80	High ($-10^6/-10^6$)

^aIn iterations in which multiple detector settings proved equally good, one was randomly chosen as top-ranked.

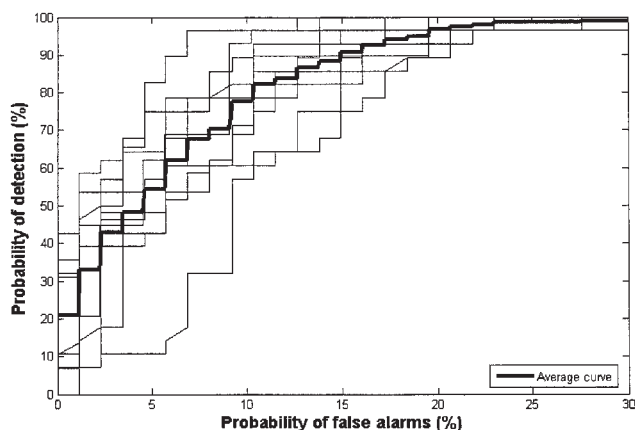


Fig. 4. Detector estimates for external CV. The graph shows the 10 ROC curves for each allergen detector i , designed with Allergen Master set i and Non-Allergen Design set i , and validated with the Allergen and Non-Allergen Validation sets i . For each allergen detector i , the detector setting selected from the internal CV is used. The thicker line shows the average detection for all 10 curves at a given level of false alarms. (Note that the probability of false alarm axis is between 0–30%.)

validation. Together with the aforementioned work of Bhasin and Raghava,³⁷ this is one of very few reports we have been able to find on protein detection and classification that clearly describes implementation of an unbiased external test procedure.

In this article, we suggest an alternative procedure (based on an external CV loop), which also is aimed at separating the selection of detector setting from performance estimation. Our results on simulated data clearly show that the estimate obtained after selection among several alternatives is biased. The results also demonstrate the feasibility of an external CV loop to obtain unbiased performance estimates. The advantage of the proposed external CV loop, compared with a single external (independent) holdout set, can be segmented into two principally distinct categories:

1. The external CV loop uses the available examples more efficiently in the sense that more examples can be used

for detector design, which is a well-known advantage and motivation for CV in comparison with holdout testing.

2. The external CV loop yields statistics on how often different parameter settings are selected (see Table II).

Detector settings occurring in the majority of the i external iterations support the assumption that these parameters are favorable for the construction of the detector in general and not solely for examples used in the internal validation. The ability to detect such tendencies (or lack of tendencies) is an important advantage over a single holdout validation test. We would like to stress that a single external holdout test results in an unbiased estimate of the finally selected detector, whereas in the procedure presented here, the final estimate reflects the average performance for the detector design procedure used. Thus, the external CV loop enables the analysis of the robustness of the detection system in parameter selection as well as in performance estimation even though no final detector is specifically evaluated. Moreover, the internal CV loop for parameter selection used in this work could be replaced by other methods such as a single holdout test or any bootstrap method.

The sample sizes in the simulations were rather small compared with the specific case of the application to detection of potential allergens, but they were chosen to make the simulations fast and illustrative. Moreover, in many biological prediction systems, the amount of data is scarce, which supports the relevance of using small samples in the simulation. With larger datasets, the bias of interest in this specific computer simulation would be less pronounced and therefore more difficult to visualize. A more complex detection problem (in this case only one parameter is optimized) would probably increase the bias, but because our purpose is to demonstrate the existence of the bias, a simplistic experimental outline was preferred. It should also be noted that small datasets increase the uncertainty in the obtained estimates, regardless of whether holdout, CV, or bootstrap methods are used

(either internally for parameter optimization or externally for unbiased error rate estimation).

A seemingly specific but quite general problem associated with amino acid sequence detection pertains to identification of a small set of reliable features of the sequences that yields powerful discrimination. This is an important issue because most supervised learning algorithms rely on feature vectors of fixed length (with few exceptions such as hidden Markov models). One successful and frequently used approach, proposed by Rost and Sander⁴⁴ for protein structure prediction, is to perform a multiple alignment for every unknown sequence, and subsequently represent that sequence in terms of features that characterizes its multiple alignment with the closest neighbors in sequence space. As regards the detection (two-class) problem considered in this work, a slightly different approach is used in which features are extracted using individual alignments of the unknown sequences to a diverse set of sequences that are known to belong to the class of interest (allergens). For both methods, several parameters have to be selected. (The choice of prototype number, substitution matrix, and gap penalty setting can be viewed as feature selection parameters because they directly influence the features ASs and ALs.) As elaborated in this work, all such parameters must be selected inside an internal validation procedure to avoid introduction of an undesired bias.

Why Are Certain Settings Preferred in the Allergen Detector Design?

Results, from the selection of settings for detecting allergen potential, listed in Table II, indicate a preference for kNN over the two Gaussian counterparts in all the external CV iterations. There is also a consistency on choice of substitution matrix as well as gap penalty setting, as revealed by a preference for Blosum80 and high gap penalty in all CV iterations. This robust pattern of supervised algorithm and parameter settings is indicative of their relevance to the construction of detectors for allergenic potential. Our first work on detection of allergenic potential²⁵ describes a detector specific for food allergens. The preferred detector algorithm used in that work was kNN with $k = 9$ as the favored amount of neighbors, largely in line with the results presented herein. In a subsequent work,²⁹ aimed at detecting any kind of allergen, extended sets of supervised algorithms and local alignment parameters were evaluated. The overall best detector in that work was based on the linear Gauss algorithm with a feature vector consisting of the mixed ASs using different substitution matrices. When a feature vector composed of AS and AL was used (as in this work), however, kNN proved to be the most robust supervised algorithm, similar to results presented herein. Regarding gap penalty settings, results reported herein show a preference for high gap opening/extension penalties, which differs from findings in the earlier study.²⁹ The consequence of using high gap penalties is that shorter ungapped regions, which may contain linear epitopes necessary for allergic reactions, become favored. The discrepancies might in part result from differences in the

datasets used but may also reflect procedural differences, because the external CV procedure described herein was not used in any of the previous work.

Markedly variable detector settings across the external CV iterations could be attributable to at least two distinct mechanisms: The most obvious reason is that the detector design procedure is not robust enough and accordingly too dependent of design data. This is plausible in cases with small data sets. Discrepancies between iterations may also occur if many settings of a single parameter are evaluated, or if there are many different parameters to select from. If the difference between two detector settings is small it is likely that their respective estimates in the internal validation procedure will be almost identical, leading to difficulties in distinguishing among the best settings. In these cases, it will be necessary to analyze not only the top-ranked setting (in each iteration) but also the settings yielding estimates close to those of the best. Otherwise a detector setting that performs well in all iterations may be overlooked, because another setting yields a slightly better estimate in one of the iterations.

General Molecular Family Detection

The problem of protein family detection may be considered as a branch of molecular family detection, i.e., detection of molecule groups of interest. A recent example of this is the detection of molecules that are permeable with respect to a particular barrier.⁴⁵ Therefore, it should be accentuated that the bias discussed in this study is not restricted to in silico detection of protein families but may appear in any molecular family detection problem wherein validation procedures are used to evaluate performance of many alternative parameter settings. This source of bias should always be taken into consideration before the design of such detectors and the procedure presented herein should be helpful in that process.

CONCLUSION

This work demonstrates a frequently overlooked bias in CV-based evaluation of in silico protein family detectors. One solution to this problem is to use an external test set and another is to embed the design procedure in an external CV loop as suggested in this work. When applied to detection of potential protein allergens, the proposed external CV loop procedure shows that particular subsets of parameter settings are selected much more often than others. This demonstrates that the suggested external CV loop is important not only for bias elimination, but also for improved understanding of classification algorithms and parameter settings to be selected for a particular application.

ACKNOWLEDGMENTS

Ms. Maj Olausson is acknowledged for expert assistance on computer graphics.

REFERENCES

1. Cserzo M, Wallin E, Simon I, von Heijne G, Elofsson A. Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method. *Protein Eng* 1997;10(6): 673–676.

2. Natt NK, Kaur H, Raghava GP. Prediction of transmembrane regions of beta-barrel proteins using ANN- and SVM-based methods. *Proteins* 2004;56(1):11–18.
3. Martelli PL, Fariselli P, Casadio R. An ENSEMBLE machine learning approach for the prediction of all-alpha membrane proteins. *Bioinformatics* 2003;19(Suppl 1):i205–211.
4. Schneider G, Fechner U. Advances in the prediction of protein targeting signals. *Proteomics* 2004;4(6):1571–1580.
5. Nielsen H, Brunak S, von Heijne G. Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng* 1999;12(1):3–9.
6. Bendtsen JD, Nielsen H, von Heijne G, Brunak S. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* 2004;340(4):783–795.
7. Nielsen H, Engelbrecht J, Brunak S, von Heijne G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng* 1997;10(1):1–6.
8. Emanuelsson O, Nielsen H, Brunak S, von Heijne G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol* 2000;300(4):1005–1016.
9. Nair R, Rost B. Better prediction of sub-cellular localization by combining evolutionary and structural information. *Proteins* 2003;53(4):917–930.
10. Bhasin M, Raghava GP. ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res* 2004;32(Web Server issue):W414–419.
11. Lu Z, Szafron D, Greiner R, et al. Predicting subcellular localization of proteins using machine-learned classifiers. *Bioinformatics* 2004;20(4):547–556.
12. Nair R, Rost B. LOCnet and LOCtarget: sub-cellular localization for structural genomics targets. *Nucleic Acids Res* 2004;32(Web Server issue):W517–521.
13. Cui Q, Jiang T, Liu B, Ma S. Esub8: a novel tool to predict protein subcellular localizations in eukaryotic organisms. *BMC Bioinformatics* 2004;5(1):66.
14. Feng ZP. An overview on predicting the subcellular location of a protein. *In Silico Biol* 2002;2(3):291–303.
15. Han LY, Cai CZ, Lo SL, Chung MC, Chen YZ. Prediction of RNA-binding proteins from primary sequence by a support vector machine approach. *RNA* 2004;10(3):355–368.
16. Zhang SW, Pan Q, Zhang HC, Zhang YL, Wang HY. Classification of protein quaternary structure with support vector machine. *Bioinformatics* 2003;19(18):2390–2396.
17. Kaur H, Raghava GP. Prediction of alpha-turns in proteins using PSI-BLAST profiles and secondary structure information. *Proteins* 2004;55(1):83–90.
18. Karchin R, Karplus K, Haussler D. Classifying G-protein coupled receptors with support vector machines. *Bioinformatics* 2002;18(1):147–159.
19. Song J, Tang H. Accurate classification of homodimeric vs other homooligomeric proteins using a new measure of information discrepancy. *J Chem Inf Comput Sci* 2004;44(4):1324–1327.
20. Cai CZ, Han LY, Ji ZL, Chen YZ. Enzyme family classification by support vector machines. *Proteins* 2004;55(1):66–76.
21. Gendel SM. The use of amino acid sequence alignments to assess potential allergenicity of proteins used in genetically modified foods. *Adv Food Nutr Res* 1998;42:45–62.
22. Gendel SM. Sequence analysis for assessing potential allergenicity. *Ann NY Acad Sci* 2002;964:87–98.
23. Hileman RE, Silvanovich A, Goodman RE, et al. Bioinformatic methods for allergenicity assessment using a comprehensive allergen database. *Int Arch Allergy Immunol* 2002;128(4):280–291.
24. Kleter GA, Peijnenburg AA. Screening of transgenic proteins expressed in transgenic food crops for the presence of short amino acid sequences identical to potential, IgE-binding linear epitopes of allergens. *BMC Struct Biol* 2002;2(1):8.
25. Zorzet A, Gustafsson M, Hammerling U. Prediction of food protein allergenicity: a bioinformatic learning systems approach. *In Silico Biol* 2002;2:525–534.
26. Ivanciuc O, Schein CH, Braun W. Data mining of sequences and 3D structures of allergenic proteins. *Bioinformatics* 2002;18(10):1358–1364.
27. Ivanciuc O, Schein CH, Braun W. SDAP: database and computational tools for allergenic proteins. *Nucleic Acids Res* 2003;31(1):359–362.
28. Stadler MB, Stadler BM. Allergenicity prediction by protein sequence. *FASEB J* 2003;17(9):1141–1143.
29. Soeria-Atmadja D, Zorzet A, Gustafsson MG, Hammerling U. Statistical evaluation of local alignment features predicting allergenicity using supervised classification algorithms. *Int Arch Allergy Immunol* 2004;133(2):101–112.
30. Li KB, Issac P, Krishnan A. Predicting allergenic proteins using wavelet transform. *Bioinformatics* 2004;20(16):2572–2578.
31. Bjorklund AK, Soeria-Atmadja D, Zorzet A, Hammerling U, Gustafsson MG. Supervised identification of allergen-representative peptides for in silico detection of potentially allergenic proteins. *Bioinformatics* 2005;21(1):39–50.
32. Cai CZ, Han LY, Ji ZL, Chen X, Chen YZ. SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res* 2003;31(13):3692–3697.
33. Keck HP, Wetter T. Functional classification of proteins using a nearest neighbour algorithm. *In Silico Biol* 2003;3(3):265–275.
34. Wu C, Whitson G, McLarty J, Ermongkonchai A, Chang TC. Protein classification artificial neural system. *Protein Sci* 1992;1(5):667–677.
35. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference and prediction. New York: Springer-Verlag; 2001.
36. Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci USA* 2002;99(10):6562–6566.
37. Bhasin M, Raghava GP. SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence. *Bioinformatics* 2004;20(3):421–423.
38. Ortiz AR, Pisabarro MT, Gago F, Wade RC. Prediction of drug binding affinities by comparative binding energy analysis. *J Med Chem* 1995;38(14):2681–2691.
39. Freyhult E, Prusis P, Lapinsh M, Wikberg JE, Moulton V, Gustafsson MG. Unbiased descriptor and parameter selection confirms the potential of proteochemometric modelling. *BMC Bioinformatics* 2005;6(1):50.
40. Vining DJ, Gladish GW. Receiver operating characteristic curves: a basic understanding. *Radiographics* 1992;12(6):1147–1154.
41. Pearson WR. Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol Biol* 2000;132:185–219.
42. Duin RPW, Juszczak P, Paclik P, Pekalska E, de Ridder D, Tax DMJ. PRTools4, A Matlab Toolbox for Pattern Recognition Research. Faculty EWI-ICT, Delft University of Technology; 2004.
43. Liu J, Rost B. Sequence-based prediction of protein domains. *Nucleic Acids Res* 2004;32(12):3522–3530.
44. Rost B, Sander C. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc Natl Acad Sci USA* 1993;90(16):7558–7562.
45. Doniger S, Hofmann T, Yeh J. Predicting CNS permeability of drug molecules: comparison of neural network and support vector machine algorithms. *J Comput Biol* 2002;9(6):849–864.