Support vector machines for protein functional classification

Jaques Reifman[1], Nela Zavaljevski[2] and Fred J. Stevens[2,3]
[1]Telemedicine and Advanced Technology Center US Army Medical Research and Materiel Command 504 Scott St., Ft. Detrick, MD 21702, USA, jaques.reifman@det.amedd.army.mil
[2]Reactor Analysis and Engineering, Biosciences Argonne National Laboratory
9700 S. Cass Ave., Argonne, IL 60439, USA, nelaz@ra.anl.gov
[3]fstevens@anl.gov

Abstract
We demonstrate that support vector machines (SVMs) with selective kernel scaling are an effective tool in discriminating between benign and pathologic proteins. Initial results compare favorably against manual classification performed by experts and indicate the capability of SVMs to capture the underlying structure of the data. The data set consists of 70 proteins of human antibody κ1 immunoglobulin light chains, each represented by aligned sequences of 120 amino acids. We perform feature selection based on a first-order adaptive scaling algorithm, which confirms the importance of changes in certain amino acid positions and identifies other positions that are key in the characterization of protein function.

Introduction
The concurrent explosions of sequence and structural genetic data have not been paralleled by an increase in the number of workers to correlate the data and extract meaningful new information and knowledge. This extraction process is critical to the way by which sequence and structure data contribute to both basic and applied research. Clearly, as neither the human genome project nor the emerging structural genomics program was possible without the introduction of extensive automation of experimental methods, a similar challenge is presented by the need to merge sequence, structure, and functional data to construct an understanding of each protein system. Such capability can be accomplished through bioinformatics tools that relate extensive amino acid variation in a protein of known structure to achieve automated prediction of a functional attribute.

As an initial step, we describe here an approach based on support vector machine (SVM) technology with selective kernel scaling for the classification of the κ family of human antibody light chains into benign or pathogenic categories and for the identification of markers, i.e., the selection of features, in the sequence of amino acids that are key discriminatory indicators. The selection of SVMs technology is driven primarily by their unique ability to construct predictive models with superior generalization power when the dimensionality of the data is high, i.e., the number of input features is large,

and the number of observations available for developing (i.e., training) the model is limited. Their selection in this study is also attributed to their property of being capable of adapting to the problem at hand by including prior knowledge into the so-called kernel (mapping) function. We make use of this property to selectively scale the importance of amino acids in the sequence based on position variability at the germline level and position discriminatory power obtained through post-processing. In this work, we employ a version of the SVMlight code (Joachims, 1999) that we have modified to include selective kernel scaling. The original code is available at http://ais.gmd.de/~thorsten/svm_light.

System and methods

Data set and encoding scheme

In this study, we employ a subset of the human antibody light chain sequences from patients with plasma cell diseases recently analyzed by Stevens (2000) in which he identified four structural "risk factors" that appear to reveal most amyloidogenic κ1 light-chains. The employed data set consists of 70 κ1 light chain proteins. Of those, six proteins were known to be benign, 33 were known to be pathogenic, i.e., from patients with myloidosis, and 31 were of unknown pathology. Further analysis of the 31 unclassified proteins, including the use of the SVM classifier itself to identify misclassified proteins, allowed us to categorize 28 proteins into the benign class and the remaining three into the pathogenic class. Therefore, the final data set is almost equally divided (34/36) between the two classes, which avoids the construction of a class-biased classifier.

The SVM classifier receives a sequence of amino acids representing a protein as its input and predicts the class of the protein as its output. Because SVMs, as well as other machine-learning algorithms, use numerical values as inputs, they require the definition of encoding schemes. The encoding scheme for protein sequences can be rather involved and can greatly impact the performance of the classifier. One possibility is to encode each one of the 20 letters corresponding to the 20 amino acid types of a protein into a numerical scheme representing known physicochemical properties of each amino acid type (Baldi et al., 1998). Here, each amino acid is represented by a set of six physicochemical properties (Lohman et al., 1994), hydrophobicity, hydrophilicity, volume, surface area, bulkiness, and refractivity, scaled to the [-1,1] interval. The primary structure of the κ light chains is aligned to 120 amino acid positions, which are, therefore, represented by 720 (120x6) input features to the SVM.

Support vector machines

Support vector machines, a recently proposed supervised machine learning technique (Vapnik, 1998), have been shown to be an effective bioinformatics tool in multiple areas of biological analysis (Zien et al., 2000; Jaakkola et al., 2000; Hua et al. 2001; Ding and Dubchak, 2001). Their unique ability to develop models with superior generalization capabilities when the number of input features is large compared to the number of training samples provides a significant advantage over other supervised learning algorithms, including neural networks (NNs). Unlike NNs where the number of model parameters that require estimation grow exponentially with the number of input features, the dimension of the SVM optimization problem is equal to the number of training samples. This unique capability affords their use for protein classification where the data sample is sparse and the dimension of the input features is large. The use of NNs, if attempted for this class of problems, would result in an overfitted model with very poor generalization capability.

When used for classification, SVMs map the input space into a higher-dimensional feature space that separates a given set of binary-labeled training data with an optimal hyperplane. The optimal hyperplane found by the SVM learning algorithm is the one that maximizes the separating margin between the binary classes of the training data and is defined by a relatively small number of M_S vectors in the input data set called support vectors. Given a training set of M samples or input vectors $\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots, \boldsymbol{x}_i,\ldots,\boldsymbol{x}_M\}$ with known class labels $\{y_1, y_2,\ldots y_i,\ldots, y_M\}$, $y_i \in \{+1,-1\}$, a new data point $\boldsymbol{x}$ is assigned a label by the SVM according to the decision function

$$f(\boldsymbol{x}) = \text{sign}\left(\sum_{i=1}^{M_S} y_i \alpha_i k(\boldsymbol{x}_i,\boldsymbol{x}) + b\right) \qquad (1)$$

where $k(\boldsymbol{x}_i, \boldsymbol{x})$ is the kernel function that defines the feature space, b is a bias value, and α_i are positive real numbers obtained by solving a quadratic programming (QP) problem that yield the maximal margin hyperplane (Vapnik, 1998).

One of the most common kernels is the polynomial kernel, $k_P(\boldsymbol{x}_i, \boldsymbol{x}_j) = (a + b\,\boldsymbol{x}_i \cdot \boldsymbol{x}_j)^{\text{“}}$, where a, b, and d are real-valued constants. A special case of this kernel is the linear kernel obtained when a=0 and b=d=1. The kernel function, however, can be customized to the problem at hand (Zien et al., 2000). This unique feature of SVMs gives us the ability to implicitly incorporate prior knowledge, such as known physicochemical protein properties, into the mapping function by properly engineering the kernel function.

Selective Kernel scaling

We customize the kernel function so that each component or input feature variables ℓ of the input vector i, x_i^{ℓ}, ℓ=1,2,…,720 and i=1,2,…,M=70, could

have a different scaling factor related to its importance to the classification problem. The modified kernel has the form $k_s(\boldsymbol{x}_i,\boldsymbol{x}_j)= k(\mathbf{S}\boldsymbol{x}_i,\mathbf{S}\boldsymbol{x}_j)$, where $\mathbf{S}$ is a diagonal matrix of scaling factors. Here, we employ equal scaling for each group of six properties representing each amino acid type, and position-dependent scaling, based on the two schemes described below, to each of the 120 positions in the amino acid sequence.

Germline scaling: The first selective kernel scaling, termed germline scaling, is based on the assessment of the significance of the variability at specific positions in the amino acid sequence. All human κ light chains originate from a repertoire of about 14 inherited or germline genes. When these sequences are compared, 40 sites are invariant at the genetic level while other positions exhibit two or more alternative amino acids. We assume that positions that are conserved at the germline level would tend to have a higher probability of significantly affecting protein fold and/or stability, and therefore, lowered the weights assigned to positions of amino acid that exhibit variability at the germline level. Accordingly, the germline scaling factor for position n, n=1,2,…,120, is computed as 1/N(n), where N(n) is the number of different amino acid types that appear at position n in the germline sequences.

Adaptive scaling: The second scheme, based on the post-processing of the classification problem, is adaptive and iteratively modifies the scaling factor of each input feature variable based on its affect or sensitivity on the classification. The sensitivity index SI_ℓ of the classification function to a change in component ℓ of input feature vector i is to the first order of approximation given by (Evgeniou, 2000)

$$SI_\ell \approx \sum_{i=1}^{M_S} \left| \frac{df}{dx_i^\ell} \right| = \sum_{i=1}^{M_S} \left| \sum_{j=1}^{M_S} \alpha_j y_j \frac{d(k(\boldsymbol{x}_i, \boldsymbol{x}_j))}{dx_i^\ell} \right|. \qquad (2)$$

To compute the scaling factor of each group of six input feature variables representing each amino acid in the sequence we add the SI in Eq. (2) over the six properties, normalize the cumulative SI to the [0,1] interval and take the square root. When germline scaling is used in conjunction with adaptive scaling, the effective scaling factor for position n is taken as the square root of the product of 1/N(n) times the normalized cumulative SI value. The scaling factor provides a measure of the sensitivity of the classifier to perturbations of each amino acid position, and therefore, it is used here as a metric for feature selection.

Simulation results

Our method is tested in a number of simulation runs with the *SVMlight* code (Joachims, 1999) modified to include the scaling kernel schemes described above. The results are compared against a manual classification approach (Stevens, 2000). Three measures of accuracy, classification error (E), recall (R), and precision (P), are used to assess the performance of the SVM

classifier for the testing data

$$E = \frac{FP + FN}{TP + FP + TN + FN} \times 100\% \qquad R = \frac{TP}{TP + FN} \times 100\%$$

$$P = \frac{TP}{TP + FP} \times 100\% \qquad (3)$$

where TP is the number of true positives, i.e., pathogenic proteins, FP is the number of false positives, TN is the number of true negatives and FN is the number of false negatives.

The classification results using the leave-one-out cross validation procedure and a linear kernel function with different position-dependent scaling schemes for the input features representing the sequence of amino acids are presented in Table 1. The results of applying germline sequence scaling based on *a priori* knowledge about the significance of each specific amino acid mutation derived from conservation at the germline level achieved 80% classification accuracy (or alternatively, 20% classification error), 80% recall accuracy, and 80% precision accuracy. The combination of the germline sequence scaling followed by adaptive scaling employing the first-order sensitivity index in Eq. (2) yields significant improvements (except in recall) ranging from 13% to 30%, confirming our hypothesis that the use of adaptive scaling results in an improved classifier.

Table 1. Leave-one-out classification accuracy based on several scaling schemes of the input features

Scaling scheme	Classification Error (E) (%)	Recall (R) (%)	Precision (P) (%)
Germline sequence scaling	20	80	80
Germline sequence scaling followed by adaptive scaling	14	80	90
Randomly assigned classes with germline sequence scaling	52	58	46
Randomly assigned classes with germline sequence scaling followed by adaptive scaling	51	38	46
Heuristic classification	15	94	79

To determine if the SVM was indeed learning the underlying structure of the data we repeated the simulations with randomly assigned labels for all 70 samples. The hypothesis being that, if the classifier was indeed learning the

underlying structure of the data, as opposed to learning the structure of random data, its accuracy with randomly assigned labels should be about 50%. A large accuracy would indicate that the SVM is learning to explain noise. The results using germline sequence scaling are illustrated in the third row of Table 1, which indicate an overall classification error of 52%, clearly showing that the SVM is capable of capturing the underlying structure of the amino acid sequences.

These simulations also serve to verify that the proposed adaptive scaling kernel algorithm does not result in an overfitted model that improves the explanation of the training data alone without improving the generalization of the classification model. When adaptive scaling was employed to the samples with randomly assigned labels the classification error remained essentially unchanged while recall decreased, see row four in Table 1. An increase in the classification accuracy would have indicated that adaptive scaling is forcing the model to learn the structure of random data.

Figure 1 shows the normalized effective scaling factors aggregated over the entire data set for the 120 amino acid positions in the sequence. They can be employed as metrics for feature selection as they provide a measure of relevance of the contribution of each amino acid position to the classification. Hence, amino acid positions with large scaling factor values are key in discriminating between benign and pathogenic proteins. Indeed, each of the three amino acid positions with large effective scaling factor values, Pro40, Arg61, and Pro95, has significant structural significance. For instance, Stevens (2000) has inferred that the interaction of Arg61 with Asp82 contributes significant free energy to the stability of the protein and any substitution of Arg61 is destabilizing and strongly associated with amyloid formation.

The capability of the scaling factors to perform feature selection can be validated by retraining the SVM classifier with the three key amino acid positions removed. By doing so, the classification accuracy deteriorated, E=37%, R=63%, and P=63%, clearly indicating that the scaling factors based on the SI in Eq. (2) provide a good mechanism for feature selection.

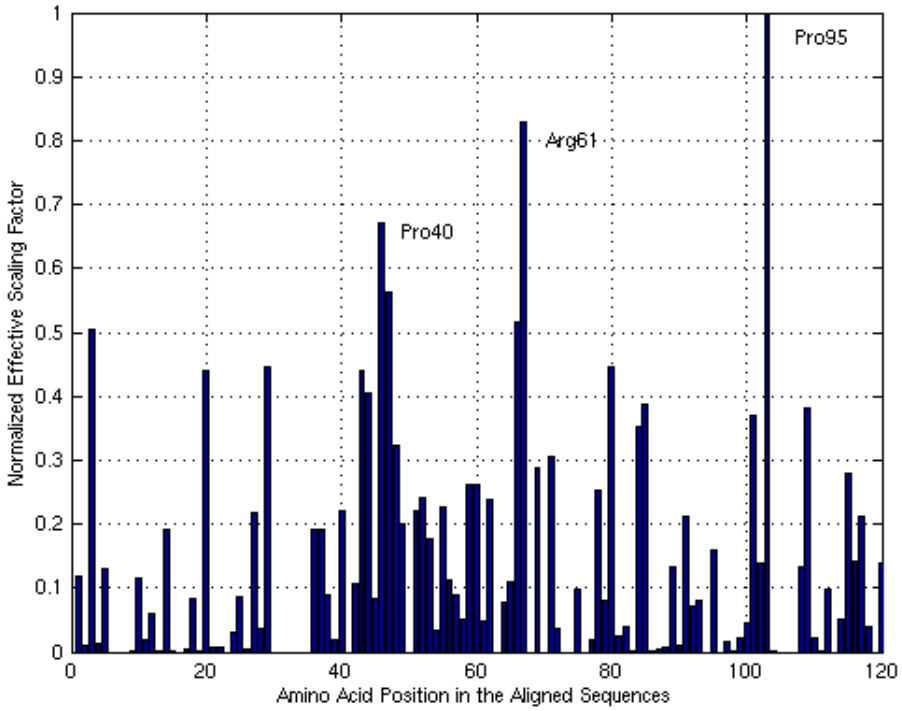

Figure 1. Positional relevance of the amino acid sequence to classification accuracy based on mutations at the germline level and first-order sensitivity index

To investigate the stability of the selective kernel scaling algorithms with regards to kernel forms we repeated the simulations using a polynomial kernel. Numerous simulations involving various changes in the three parameters (a, b, and d) of the polynomial kernel resulted in identical classification accuracy as the ones obtained with the linear kernel. Furthermore, comparisons of the effective scaling factors indicate only minimal variations around the distribution depicted in Figure 1. The importance of the same three amino acid positions, Pro40, Arg61, Pro95, was unmistakably distinguished in every simulation, which serves to demonstrate that the adaptive scaling algorithm is inherently stable, independent of kernel form.

The last row in Table 1 shows the results of a heuristic classification based on manual analysis of the data (Stevens, 2000). Because all 70 samples were used to infer the heuristic classification rules, we cannot assess the generalization ability of the approach and perform a consistent comparison with the SVM

results. Nonetheless, the heuristic approach provides for a semi-quantitative comparison, which indicates a favorable performance of the SVM algorithm.

Conclusions

Preliminary results demonstrate the ability of the Support Vector Machine algorithm with selective kernel scaling to discriminate between benign and pathological immunoglobulins. Evaluations using the leave-one-out error estimate compares favorably with the accuracy obtained through manual heuristic classification performed by domain experts. Simulation tests where the protein labels, benign and pathological, are randomly assigned to the data are used to verify that the modified SVM is capable of capturing the underlying structure of the data. SVMs provide an effective inductive tool for developing protein classification models where the data is sparse and the dimensionality of the input features is large.

The use of adaptive scaling based on first-order sensitivity analysis is shown to be a particular important method to improve the classification accuracy and allow for feature selection. It improves classification error by 30% and confirms the importance of certain amino acid positions in the light chain, such as Arg61, and identifies new amino acid positions, such as Pro40 and Pro95, which contribute significantly to determining protein stability, previously shown to be the principal determinant of the pathological attribute of light chain pathology (Raffen et al., 2000). The approach is stable in regards to kernel forms, providing the same performance improvements independent of the type of kernel used.

In future studies, we will explore new algorithms for scaling the input feature variables and performing feature selection. In addition, we will investigate ways in which the information content of the three-dimensional protein structure can be implicitly embedded in the scaling procedure. We believe that it is imperative to combine protein primary structure information with tertiary structure information to characterize the protein functional behavior—a critical feature ignored by simple analysis of strings of amino acid labels.

Acknowledgements

The authors want to express their gratitude to T. Joachims for providing access to the SVMlight code. The first author was supported in part by the Combat Casualty Care and Military Operational Medicine research programs of the U.S. Army Medical Research and Materiel Command. The last author was supported by the U.S. Department of Energy, Office of Biological and Environmental Research, under contract W-31-109-ENG-38 and by USPHS Grants DK43757 and AG1001.

References

Baldi,P. and Brunak,S. (1998) *Bioinformatics - The Machine Learning Approach*. MIT Press, Cambridge.

Carrell,R.W. and Gooptu, B. (1998) Conformational changes and disease – serpins, prions, and Alzheimer's. Curr. Opin. Struct. Biol., 8, 799-809.

Ding,C.H.Q. and Dubchak,I. (2001) Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics, 17, 349-358.

Evgeniou,T., Pontil,M., Papageorgiou,C. and Poggio,T. (2000) Image representation for object detection using kernel classifiers, 4th Asian Conference on Computer Vision, January 9-11, Taipei, Taiwan, Paper ACCV-198, Poster Session.

Hua,S. and Su,Z. (2001) A novel method of protein secondary structure prediction with segment overlap measure: Support vector machine approach. Journal of Molecular Biology, in press.

Jaakkola,T., Diekhans,M. and Haussler,D. (2000) A discriminative framework for detecting remote protein homologies. Journal of Computational Biology, 17, 95-114.

Joachims,T. (1999) Making large scale SVM learning practical. Advances in Kernel Methods - Support Vector Learning, B. Scholkopf, C. Burges, and A. Smola (ed.). MIT Press, Cambridge.

Lohman,R., Schneider,G., Nehrens,D. and Wrede,P. (1994) A neural network model for the prediction of membrane-spanning amino acid sequences. Protein Science, 3, 1597-1601.

Raffen,R., Dieckman,L.J., Szpumar,M., Wunschl,C. Pokkuluri,P.R., Dave,P., Wilkins,S.P., Cai,X., Schiffer,M., and Stevens,F.J. (1999) Physicochemical consequences of amino acid variations that contribute to fibril formation by immunoglobulin light chains. Protein Science, 8, 509-17.

Stevens,F.J. (2000) Four structural risk factors identify most fibril-forming kappa light chains. Amyloid: Int. J. Exp. Clin. Invest., 7, 200-211.

Vapnik,V. (1998) Statistical Learning Theory. Wiley, NY.

Zien,A., Ratsch,G., Mika,S., Scholkopf,B., Lengauer,T. and Muller,K.R. (2000) Engineering Support Vector Machine kernels that recognize translation initiation sites, Bioinformatics, 16, 815-824.