Sparse Logistic Classifiers for Interpretable Protein Homology Detection

Pai-Hsi Huang, Vladimir Pavlovic
Dept. of Computer Science, Rutgers University
Piscataway, NJ 08854-8019 U.S.A

{paihuang,vladimir}@cs.rutgers.edu

Abstract

Computational classification of proteins using methods such as string kernels and Fisher-SVM has demonstrated great success. However, the resulting models do not offer an immediate interpretation of the underlying biological mechanisms. In particular, some recent studies have postulated the existence of a small subset of positions and residues in protein sequences may be sufficient to discriminate among different protein classes. In this work, we propose a hybrid *setting for the classification task. A generative model is trained as a feature extractor, followed by a* sparse *classifier in the extracted feature space to determine the membership of the sequence, while discovering features relevant for classification. The set of* sparse *biologically motivated features together with the discriminative method offer the desired biological interpretability. We apply the proposed method to a widely used dataset and show that the performance of our models is comparable to that of the state-of-the-art methods. The resulting models use* fewer than 10% *of the original features. At the same time, the sets of critical features discovered by the model appear to be consistent with confirmed biological findings.*

1. Introduction

Protein homology detection is a fundamental problem in computational biology. With the advance of large-scale sequencing techniques, it becomes evident that experimentally determining the function of an unknown protein sequence is an expensive and tedious task. Currently, there are more than 54 million DNA sequences in GenBank [3], and approximately 208,000 annotated and 2.6 million unannotated sequences in UNIPROT [1] . The rapid growth of sequence databases makes development of computational aids for functional annotation a critical and timely task.

Early approaches to computationally-aided homology detection, such as BLAST [22] and FASTA [23], rely on aligning the query sequence to a database of known sequences (pairwise alignment). However, alignment is performed on the query sequence to *each* of the sequences in the database *one at a time*. Later methods, such as profiles [7] and profile hidden Markov models (profile HMM) [5] collect aggregate statistics from a group of sequences known to belong to the same family. Upon query time, an unknown sequence is aligned to all models to test for significant *hits*. Profile HMMs have demonstrated great success in protein homology detection. The linear structure of a profile HMM offers great interpretability to the underlying process that generates the sequences: the *match* states represent positions in the superfamily that are *conserved* throughout the evolutionary process. However, as *generative* models, profile HMMs are estimated from sequences known to belong to the same superfamily and do not attempt to capture the differences between members and non-members. Also, it has been shown that profile HMMs are unable to detect members with low sequence identity.

To tackle these deficiencies, Jaakkola *et al.* proposed *SVM-Fisher* in [10]. The idea is to combine a generative model (profile HMM) with a discriminative model (support vector machines, SVM) and perform homology detection in two stages. In the first stage, the generative model, *trained with positive sequences only*, extracts fixed-length features from all sequences (*positive and negative*). In the second stage, given the features, the discriminative model constructs the decision boundary between the two classes.

The class of *string kernels*, on the other hand, bypasses the first stage and directly model the decision boundary using SVMs. The *spectrum kernel* [15], the *mismatch kernel* [16] and the *profile kernel* [14] define different notions of *neighborhood* for a subsequence of size $k \geq 1$ and determine the similarity between the two sequences as a function of the size of the intersection of their neighborhood.

Previous studies suggest that both approaches are more effective than the generative models. Despite their great success, these two approaches are not readily interpretable or, when an interpretation of the models is available, it may not be biologically intuitive. For instance, the model should be able to explain how sequences in the same superfamily

Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06)
0-7695-2702-7/06 $20.00 © 2006 IEEE

evolve over time. Are there certain positions that are *critical* to a superfamily? If so, what kind of physical/chemical properties should such positions possess? Although profile HMMs attempt to offer such explanations, as generative models they lack the discriminative interpretability.

The central idea of our work is to develop an interpretable method for protein homology detection. Our approach is motivated by the results presented in [12, 19, 13] that postulate the existence of a small subset of positions and residues in protein sequences may be sufficient to discriminate among different protein classes. We aim to recover these *critical positions* and the type of residues that must occur at these positions using a new set of features embedded in a class of discriminative models. The combination of the features and the classifier may offer a *simple and intuitive* interpretation to the underlying biological mechanism that generates the biosequences.

2 Related works

In [11, 10] Jaakkola *et al.* proposed to use the gradient of the log-likelihood of the sequence with respect to the model parameters as features: $f_{\tilde{x},\tilde{s}} = \xi(\tilde{x},\tilde{s})/\theta_{\tilde{x}|\tilde{s}} - \xi(\tilde{s})$ where $\tilde{\mathrm{x}} \in \Sigma$, the alphabet set, $\tilde{\mathrm{s}} \in S$, the emitting states in the model, Θ represents the set of parameters of the model, and $\xi(\tilde{x},\tilde{s})$ as well as $\xi(\tilde{s})$ are the sufficient statistics, as defined in [18]. The extracted fixed-length features are called the *Fisher scores* and used to build the SVM for homology detection. SVM-Fisher approach has received some criticism because an inference procedure of quadratic complexity is required for each sequence. Although the criticism does address a valid concern for a general HMM, in the case of a profile HMM, such issue does not exist: the linear structure enables one to make inference in linear time.

The methods based on *string kernels*, on the other hand, bypass the need of a generative model as a feature extractor. Given a sequence, X, the *spectrum-k* kernel [15] first *implicitly* maps it to a d-dimensional vector, where $d = |\Sigma|^k$. The representation of the sequence X in the feature space is $\Phi_k(X) = \sum_\alpha (I(\alpha = \gamma))_{\gamma \in \Sigma^k}$, where α denotes all *k-mers* in X and γ denotes a member in the set of all *k-mers* induced by Σ. The similarity between two sequences, X and Y, is then defined as $K(X,Y) = \Phi_k(X)^T \Phi_k(Y)$, The $mismatch(k,m)$ kernel [16] relaxes exact string matching by allowing up to m mismatches between α and γ. In such setting, each element in the kernel matrix takes $O(k^{m+1}|\Sigma|^m(T_X + T_Y))$ time to compute.

3 Proposed features and methods

Our computational approach to remote homology detection has two steps: feature extraction followed by joint classification and feature selection in the constructed feature space. A crucial aspect of this approach lies in the ability to impose the sparsity constraint, leading to significant reduction in the number of utilized features and the interpretability of the final model. We show the proposed *hybrid* procedure in Fig. 1.

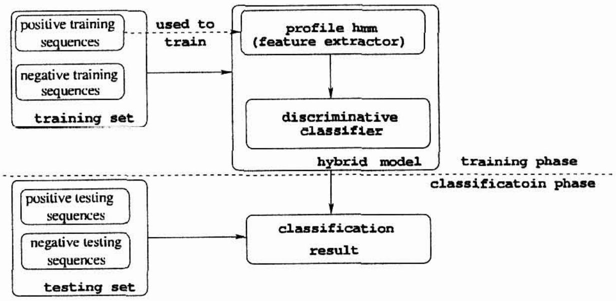

positive training sequences
negative training sequences
training set
used to train
profile hmm (feature extractor)
discriminative classifier
hybrid model
training phase
classificatoin phase
positive testing sequences
negative testing sequences
testing set
classification result

Figure 1. The proposed hybrid model.

3.1 Feature extraction

We use the sufficient statistics corresponding to the symbols of the *match* states as features. This choice of features may allow immediate biological interpretation of the constructed model because the structure of a profile HMM indicates that the *match* states represent the positions that are conserved throughout evolution. The proposed features can be obtained using the forward and backward algorithm described in [18]. In this setting, each example is represented by a vector of length $d = m|\Sigma|$, where m is the number of match states in the profile HMM and $|\Sigma| = 20$.

3.2 Classification and Feature Selection via Logistic Regression

Let f_i be the features extracted from the i^{th} example, X_i, and $c_i \in \{0,1\}$ be the response variable, where $c_i = 1$ denotes membership of the superfamily. The logistic regression model defines the probability of sequence X_i belonging to the superfamily of interest as $\pi_i = P(c(X_i) = 1) = \phi(\beta^T f_i)$, where β is the parameter of the model and $\phi(.)$ is the *cumulative distribution function* (CDF) of a logistic distribution. To estimate the model, one sets $\hat{\beta}$ to β^*, the parameter vector maximizing the joint likelihood of the observed data. There are existing algorithms for estimating β, such as *Iteratively Reweighted Least Squares* algorithm. Like SVM, the logistic model is a *discriminative* classifier.

3.3 Interpretation of the logistic model with the proposed features

Use of the logistic model provides a simple and intuitive description of data. If the assumption, $p(c = 1|\mathbf{f},\beta) = \phi(\mathbf{f}^T\beta)$, holds, then the contribution of each predictor variable, $f^{(j)}$, $1 \leq j \leq d$, is reflected in the corresponding model parameter, β_j. A coefficient with a large absolute value indicates a strong preference for a type of amino acids at the corresponding position: the position prefers a specific amino acid to be present when the coefficient is large and positive and prefers a specific amino acid to be absent when the coefficient is large and negative.

Moreover, β also offers a probabilistic interpretation. Define the *odds* of an event with probability p of occurring as $\frac{p}{1-p}$; given the estimated parameter $\hat{\beta}$, and a feature vector f_i representing sequence X_i in the *feature space*, the estimated *odds* of sequence X_i belonging in the superfamily can be expressed as $odds(X_i \in \ supFam) = e^{\hat{\beta}^T f_i}$. Define a new sequence $X_{i'}$ such that $f_{i'} = f_i$, except $f_{i'}^{(j)} = f_i^{(j)} + 1$, meaning we increase the j^{th} covariate of example i, by one unit. In this case, $odds(X_{i'} \in supFam) = e^{\hat{\beta}^T f_i + \hat{\beta}_j}$, indicating that the odds are multiplied by $e^{\hat{\beta}_j}$ in response to such increase. For example, suppose at position $\tilde{s}$, the parameter $\hat{\beta}^{\tilde{s},\tilde{x}}$ for symbol $\tilde{x}$ is $0.1615 = log(1.175)$. Then the odds of a sequence, X, being in the superfamily increases by 17.5 percent if in X, the symbol $\tilde{x}$ aligns to the model at position $\tilde{s}$.

One may argue that the preference for presence or absence of a specific amino acid at a position in a group of sequences is already reflected in the profile HMM and using a logistic model to recover the desired information is redundant. However, this need not be the case: one position in a specific superfamily may prefer a certain group of amino acids which is also preferred by another group of sequences. In this case the corresponding coefficient in the logistic model we proposed will be *insignificant*, close to 0. A coefficient corresponding to a certain type of amino acids at one position will be significant if, for example, it has been observed that the group of amino acids are present in the family of interest (the positive examples) and are absent in all the other families (the negative examples).

3.4 Use of Sparsity-enforcing Regularizers

Our belief that the model may be *sparse* leads us to set the prior distribution $\beta \sim N(0, A)$, where A is some covariance matrix. In our study, we set A to be some *diagonal* matrix. Such an assignment states that all features are *mutually independent*. The assumption is clearly not valid, since the features are the sufficient statistics. However, it is impractical to assume a general covariance structure as it involves either specifing or estimating $\binom{d}{2}$ parameters in advance. Also, Gaussian priors often do not set the coefficients corresponding to the irrelevant features to 0, because the shape of the distribution is too mild around the origin. Therefore, we use priors that *promote* and *enforce* sparsity: the Laplacian priors. In such setting, $P(\beta_i) = \frac{\sqrt{\gamma}}{2} e^{-\sqrt{\gamma}|\beta_i|}$. The Laplacian priors produce sparser models than Gaussian priors.

3.5 A Similar Setting with SVM

Given the feature vectors, one may choose to also build the decision boundary using an SVM. In the case of a linear kernel, like the logistic model, the SVM also builds a linear decision boundary to discriminate between the classes. However, the results produced by an SVM may be interpretable *only* when a *linear* (or possibly *polynomial*) kernel is used. While the objective functions in the SVM and logistic regression settings are different, the results are often similar.

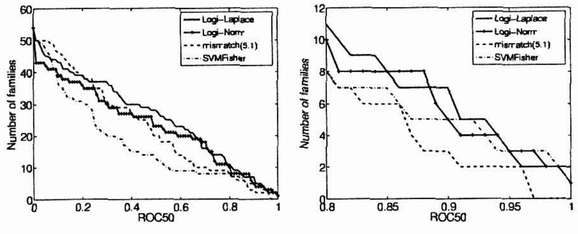
4 Experiments and Results

We use the dataset published in [25] to perform our experiments. The dataset contains 54 target families from SCOP 1.59 [17] with 7329 SCOP domains. No sequence shares more than 95% identity with any other sequence in this dataset. The dataset appears to be very diverse, where in some family, there are as few as 2 positive training sequences while in some other families, there are as many as 95 positive training sequences. Such diversity makes homology detection a challenging task. We evaluate all methods using the *Receiver Operating Characteristic* (ROC) and ROC-50 [8] scores. The ROC-50 score is the (normalized) area under the ROC curve up to 50 false positives. With small number of positive testing sequences and large number of negative testing sequences, the ROC-50 score is more indicative of the prediction accuracy.

We build all profile HMMs for our *hybrid* procedure in the following way: first, we locate the profile most suitable for the experiment and download the multiple alignment from PFam [2] and estimate an initial profile HMM from the multiple alignment; next, we refine the profile HMM with the positive training sequences using a procedure similar to the EM [4] algorithm with 9-component mixture of Dirichlet priors [21]. To avoid over-representation of sequences, we also use *position-based* weighting scheme [9].

For logistic models, we perform our experiments on Normal and Laplace priors using *Bayesian Binary Regression Software* (BBR) [6]. Precision γ in the Laplace models are set to the value suggested by [6]. Experiments using linear kernel SVM make use of an existing machine-learning package called *Spider* (available at *http://www.kyb.tuebingen.mpg.de/bs/people/spider*).

In Fig. 2, we compare the performance of the models. The figures indicate that, with ROC-50 score greater than 0.4, both logistic models dominate the mismatch kernel. The performance of both logistic models appears to be comparable in the area of high ROC-50 score (> 0.8); in the area of low scores, the logistic model with Normal prior shows slightly higher prediction accuracy. Finally, *SVM-Fisher* performs well in the area of high ROC-50 score, but the performance starts to degrade when ROC-50 score falls under 0.8. To compare the logisitic model with different



(a) (b)

Figure 2. Comparison of performance of mismatch(5,1) kernel, SVM-Fisher, and logistic model with Normal and Laplacian priors. Panel (a) shows the number of families whose ROC-50 scores are better than a given threshold. Panel (b) shows the detail plot of the high ROC-50 score region of (a).

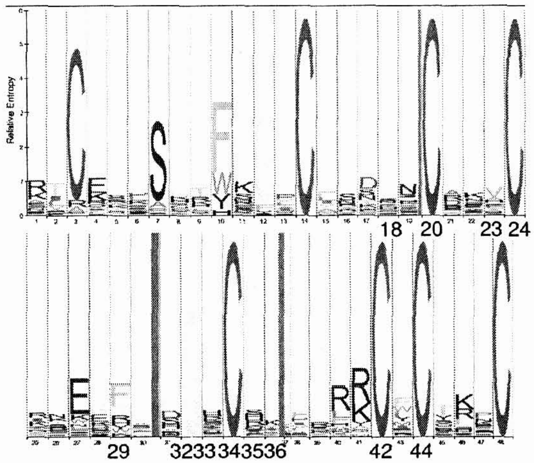


Figure 3. The HMM-logo of the *plant defensins* **family, obtained from PFam. The positions which are deemed as important by our logistic model with Laplacian prior are highlighted.**

priors, we run a sign test. The resulting p-value is close to 1, suggesting that the performance of the model with both priors are comparable.

4.1 The Sparse Model

Enforcing sparsity in the parameters can be viewed as a *feature selection* process. The logistic model with Laplacian prior discards the irrelevant features by setting the corresponding parameters to 0. Among 54 families, there are, on average, 480 features to select from. The Laplacian prior selects only about 43 features per family, resulting in more than 90% reduction in the final number of selected features.

The set of features selected by the sparse model can offer interesting insights into the biological significance of the discovered "critical positions". For example, our results indicate that the performance is consistently good on the *Scorpion toxin-like* superfamily. In one particular family, *Plant defensins*, out of 940 features (47 positions), the Laplacian prior selects 19 features, scattered on about 12 positions. The ROC-50 score of the classifier on this family is 1. We further extract these *critical positions* along with their preferred symbols: {(18[18],**E**), (20[20],**C**), (23[23],**H**), (24[24],**C**), (29[29],**G**), (34[32],**G**), (35[33],**K/Y**), (36[34],**C**), (37[35],**D/Y**), (38[36],**G/N**), (41[42],**C**), (43[44],**C**)}, where in each pair, the leading number corresponds to the position in our profile HMM, the number in the bracket corresponds to the position in the HMM-logo [20] in Fig. 3, and the letter the preferred symbol. The positions slightly disagree because we use a different heuristic to label columns in multiple alignment as *match* or an *insertion* states. It appears from the figure that our classifier captures some of the conserved cysteine residues: (20, 24, 34, 42 and 44) that are *critical for discerning* plant defensins from other similar proteins. Other conserved cysteins are not selected as critical features. Upon detailed examination, it became clear that while conserved in Plant defensines, these positions also appeared to be conserved in other similar families with the same symbol and were, therefore, not deemed discriminative. A set of 9 different residues seemed to play a more critical classification role.

Finally, it is worth noting that the dense model with Normal priors also achieved similar performance to the sparse model, but the weights learned by the dense model did not allow any immediate interpretation of importance nor does it select a small set of critical discriminative features.

4.2 Discussion

Tsuda *et al.* [24] proposed a class of *marginalized kernels* for biological sequences. It can be shown that in a kernel setting, the kernel induced by our feature set is a *marginalized count kernel*. Moreover, the kernel induced by our feature set also has a close relationship with the Spectrum kernel [15]: while our kernel is similar to a spectrum-1 kernel, the use of the feature extractor introduces additional *positional information* not present in the spectrum kernel.

5 Conclusion

In this paper we introduce a method for learning *sparse* feature models for the homology detection problem. We use a profile HMM representing the family of interest to extract the features: the sufficient statistics of the query sequence with respect to the profile HMM. As such, the features offer insight to the underlying evolutionary process such as the degree of conservation of each position in the superfamily.

Using interpretable logistic classifiers with Laplace priors, the learned models exhibit more than 90% reduction in the number of selected features. These results indicate that it may be possible to discover very *sparse* models for certain protein superfamilies, which might confirm the hypotheses suggested in [12, 19, 13] that a small subset of positions and residues in protein sequences may be sufficient to discriminate among different protein classes. We show that the *sparse* model select some *critical positions* that are consistent with current reports.

In our future work we will further investigate and consider biological interpretation of the resulting *sparse* models. At present, the full set of selected positions may not fully agree with the proposed hypotheses. Further analysis is needed to study the correspondences between the computation and hypothesized models. In addition, we will expand our framework to utilize additional sets of physically motivated features and the unlabeled data, leveraging the benefits of large training sets.

6 Acknowledgments

This work was done through graduate student support through DIMACS under NSF grants CCF-04-32013 and EIA-02-05116 to Rutgers University.

References

[1] A. Bairoch, R. Apweiler, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L.-S. L. Yeh. The Universal Protein Resource (UniProt). *Nucl. Acids Res.*, 33(suppl-1):D154–159, 2005.

[2] A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. L. Sonnhammer, D. J. Studholme, C. Yeats, and S. R. Eddy. The Pfam protein families database. *Nucleic Acids Research*, 32(Database-Issue):138–141, 2004.

[3] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. Genbank. *Nucl. Acids Res.*, 33(suppl-1):D34–38, 2005.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Jorunal of the Royal Statistical Society. Series B*, 1977(1):1–38, 1977.

[5] S. Eddy. Profile hidden Markov models. *Bioinformatics*, 14(9):755–763, 1998.

[6] A. Genkin, D. D. Lewis, and D. Madigan. Large-Scale Bayesian Logistic Regression for Text Categorization, 2006 (to appear).

[7] M. Gribskov, A. McLachlan, and D. Eisenberg. Profile analysis: detection of distantly related proteins. *Proceedings of the National Academy of Sciences*, 84:4355–4358, 1987.

[8] M. Gribskov and N. Robinson. Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching, 1996.

[9] S. Henikoff and J. G. Henikoff. Position-based sequence weights. *J Mol Biol.*, 243(4):574–8, 11 1994.

[10] T. Jaakkola, M. Diekhans, and D. Haussler. Using the Fisher kernel method to detect remote protein homologies. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, pages 149–158. AAAI Press, 1999.

[11] T. Jaakkola, M. Diekhans, and D. Haussler. A discriminative framework for detecting remote protein homologies. In *Journal of Computational Biology*, volume 7, pages 95–114, 2000.

[12] A. E. Kister, A. V. Finkelstein, and I. M. Gelfand. Common features in structures and sequences of sandwich-like proteins. *PNAS*, 99(22):14137–14141, 2002.

[13] A. E. Kister, M. A. Roytberg, C. Chothia, J. M. Vasiliev, and I. M. Gelfand. The sequence determinants of cadherin molecules. *Protein Sci*, 10(9):1801–1810, 2001.

[14] R. Kuang, E. Ie, K. Wang, K. Wang, M. Siddiqi, Y. Freund, and C. S. Leslie. Profile-based string kernels for remote homology detection and motif extraction. In *CSB*, pages 152–160, 2004.

[15] C. S. Leslie, E. Eskin, and W. S. Noble. The spectrum kernel: A string kernel for svm protein classification. In *Pacific Symposium on Biocomputing*, pages 566–575, 2002.

[16] C. S. Leslie, E. Eskin, J. Weston, and W. S. Noble. Mismatch string kernels for svm protein classification. In *NIPS*, pages 1417–1424, 2002.

[17] L. Lo Conte, B. Ailey, T. Hubbard, S. Brenner, A. Murzin, and C. Chothia. SCOP: a structural classification of proteins database. *Nucleic Acids Res.*, 28:257–259, 2000.

[18] L. R. Rabiner. A tutorial on hidden Markov models and selected apllications in speech recognition. In A. Waibel and K.-F. Lee, editors, *Readings in Speech Recognition*, pages 267–296. Kaufmann, San Mateo, CA, 1990.

[19] B. Reva, A. Kister, S. Topiol, and I. Gelfand. Determining the roles of different chain fragments in recognition of immunoglobulin fold. *Protein Eng.*, 15(1):13–19, 2002.

[20] B. Schuster-Bockler, J. Schultz, and S. Rahmann. Hmm logos for visualization of protein families. *BMC Bioinformatics*, 5(1):7, 2004.

[21] K. Sjolander, K. Karplus, M. Brown, R. Hughey, A. Krogh, I. Mian, and D. Haussler. Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Comput. Appl. Biosci.*, 12(4):327–345, 1996.

[22] S. ALTSCHUL, W. GISH, W. MILLER, E. MYERS, and D. LIPMAN. Basic Local Alignment Search Tool. *Journal of Molecular Biology*, pages 403–410, 1990.

[23] W. PEARSON and D. LIPMAN. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, 85:2444–2448, 1988.

[24] K. Tsuda, T. Kin, and K. Asai. Marginalized kernels for biological sequences. *Bioinformatics*, 18(suppl_1):S268–275, 2002.

[25] J. Weston, C. Leslie, E. Ie, D. Zhou, A. Elisseeff, and W. S. Noble. Semi-supervised protein classification using cluster kernels. *Bioinformatics*, 21(15):3241–3247, 2005.