FORMATICS

Vol. 21 Suppl. 1 2005, pages i495–i501
doi:10.1093/bioinformatics/bti1024

Classifying noisy protein sequence data: a case study of immunoglobulin light chains

Chenggang Yu[1,3,*], Nela Zavaljevski[1,3], Fred J. Stevens[1], Kelly Yackovich[2] and Jaques Reifman[3]

[1]Argonne National Laboratory, 9700 S. Cass Avenue, Argonne, IL 60439, USA, [2]Department of Computer Information Science, Clarion University of Pennsylvania, Clarion, PA 16214, USA and [3]US Army Medical Research and Materiel Command, 504 Scott Street, Fort Detrick, MD 21702, USA

Received on January 15, 2005; accepted on March 27, 2005

e classification of protein sequences obtained
with various immunoglobulin-related conform-
es may provide insight into structural correl-
enicity. However, clinical data are very sparse
se of antibody-related proteins, the collected
ve large variability with only a small subset
elevant to the protein pathogenicity (function).
these sequences represent a model system
nt of strategies to recognize the small sub-
-determining variations among the much larger
nary structure diversifications introduced during
er such conditions, most protein classification
e limited accuracy. To address this problem, we
port vector machine (SVM)-based classifier that
uence and 3D structural averaging information.
cid in the sequence is represented by a set of
mical properties: hydrophobicity, hydrophilicity,
e area, bulkiness and refractivity. Each position
ce is described by the properties of the amino
sition and the properties of its neighbors in 3D
e sequence. A structure template is selected to
ghbors in 3D space and a window size is used
he neighbors in the sequence. The test data
proteins of human antibody immunoglobulin
ach represented by aligned sequences of 120
he methodology is applied to the classification
uences collected from patients with and without

1 INTRODUCTION

Critical information relating amino acid changes with the spectrum of functional attributes exhibited by a protein is usually buried among sequence mutations irrelevant for investigated attributes. Immunoglobulin-type beta-domains, which are found in approximately 400 functional distinct forms in humans alone, provide the immense genetic variation within limited conformational changes. A protein database compiled from patients with and without amyloidosis provides unique features to serve as a model system, not only for conformational disease studies but also for the development of computational methods for analysis of structure–function relationships among evolutionarily related families. We are developing computational tools based on the support vector machine (SVM) (Vapnik, 1998) algorithm to classify proteins into pathogenic and benign classes and to identify amino acid variations that contribute to the functional attribute of pathogenic self-assembly in some human antibody light chains produced by patients with amyloidosis.

SVMs have been used recently in a wide variety of applications in computational biology (Noble, 2004). Most applications of the SVM algorithm for protein classification are based on sequence information alone (Jaakkola *et al.*, 2000; Hua and Sun, 2001; Leslie *et al.*, 2002; Cai *et al.*, 2003), as protein structures are usually unknown. Earlier, we developed an iterative SVM-based algorithm for immunoglobulin light chain classification based on protein sequence information (Zavaljevski *et al.*, 2002), where each amino acid in the

ained by the absence of significant single point
his family and/or by a higher degree of sequence
in the available data.
of some proteins to amyloid formation could
zed by specific sequence motifs, as recently
n some experimental studies (Lopez de la Paz
2004). In addition, more genetic variability is
g the λ light chains than among the κ light
ms *et al.*, 1996). To enable the analysis of mul-
ive mutations and account for the high degree of
ability, we perform classification based on pos-
orhoods where both sequential and structural
considered, separately.
sumptions are made in considering structural
s. Although there are a large number of immun-
ctures in the PDB, the vast majority of them
not humans—and the detailed structural neigh-
not known for most of the light chains in
However, since immunoglobulin light chains
r 3D structure, we assume that the structure of
n can be used for the classification of closely
chains. We anticipate that classification could
in the future, by combining information from
namics simulations with that of experimentally
ructures to infer structural information that is
each sequence.

ACH

ts

tabase of human light chain sequences from
with and without amyloidosis. Many of
es are reported in a previous paper (Stevens
and others are available in flatfiles at ftp://
s.anl.gov/VL-Database/. The database includes
ene families encoded on separate chromosomes
substantial amino acid variation. The κ family
by four major subgroups, of which the κ_1 sub-
most common. The λ family is represented by
, of which three subgroups are analyzed in this
quences are manually aligned to 120 positions,
count conserved positions in immunoglobulin
uctures. The variability of the sequences in the
set can be quantified by similarity scores based

Table 1. Data similarity scores (mean values)

Subgroup	λ_1	λ_2	λ_3	κ_1
Size	28/21[a]	19/20	20/31	36/34
$S(b, b)$	363(173)[b]	427(65)	376(90)	474(43)
$S(p, p)$	419(116)	416(90)	402(90)	467(31)
$S(b, p)$	385(154)	416(81)	387(93)	468(37)

[a] The number of sequences in the pathogenic and the benign classes.

[b] The number in parenthesis represents the standard deviation of the score.

the κ family. Second, for each family and each subgroup, there are negligible differences between the average intraclass similarity scores, $S(b, b)$ and $S(p, p)$, and the interclass similarity scores, $S(b, p)$, which represents a problem for sequence encoding based on the amino acid alphabet alone. This implies that a successful classifier ought to use additional information, such as that contained in 3D and sequence structural neighborhoods, so that the encoding (i.e. the weight) of each residue in the sequence is based not only on the amino acid type but on its position in the sequence.

2.2 SVM encoding strategy

Since experimental studies have indicated significant correlation between protein physicochemical and structural properties and protein structural stability (Gromiha *et al.*, 1999; Raffen *et al.*, 1999), we implement sequence encoding based on six physicochemical properties: hydrophobicity, hydrophilicity, volume, surface area, bulkiness and refractivity (Lohman *et al.*, 1994). This type of encoding, therefore, provides additional information for amyloid and benign protein discrimination.

Hence, the encoding of the protein sequence into the SVM algorithm is represented by a real-value vector of dimensionality equal to the length of the protein sequence (120) multiplied by the number of physicochemical properties (6) used to represent each residue. This method enables the SVM kernel function to account for the physicochemical changes in the protein sequences and simplifies the incorporation of the neighborhood information in the SVM algorithm. It is important to point out, however, that while the selected set of physicochemical properties used here was proven to be successful in our previous work (Zavaljevski *et al.*, such as the linear kernel (LK), the Gaussian polynomial kernel, a variety of string kernels, smatch kernels (Leslie *et al.*, 2002), have been ially for protein and gene classification. The els are based on inexact-matching occurrences bsequences (*k*mers).

xtend a standard kernel that takes the inner o vectors representing two protein sequences t first average the properties of a residue and r geometrical neighbors for each residue of the then take the inner product of the two vectors property entries. This allows for an area-to-on instead of a position-to-position comparison a standard kernel. The position-to-position s the simplest representation, is able to discrim-ogenic proteins characterized by point muta-nto account environmental structural changes hood area, the area-to-area comparison should ble to discriminate amyloidogenic and non-c proteins characterized by multiple mutational fferences in the amino acid sequence.

are introduced in this paper, sequential and metric) kernels. The geometric kernel, denoted defined as

$$(\boldsymbol{x}_k, \boldsymbol{x}_m) = K(S(\boldsymbol{x}_k, T), S(\boldsymbol{x}_m, T))$$
$$= K(\boldsymbol{s}_k, \boldsymbol{s}_m), \qquad (1)$$

$\boldsymbol{x}_m$ are vectors representing two amino acid nd m respectively. S denotes a mapping from ence to the 3D structure, and T is the threshold size of the 3D neighborhood to be considered. ximum neighbor distance suggested in protein ies, i.e. $T = 8.0$ Å (Gromiha *et al.*, 1999). The e vectors $\boldsymbol{s}_k$ and $\boldsymbol{s}_m$ are represented by weighted he physicochemical properties in the geometric The average value of property j for position p is denoted by $s_{m,p,j}$ and given by

$$s_{m,p,j} = \frac{\sum_{i=1}^{n_p} x_{m,I_p(i),j}\, w_{I_p(i)}}{\sum_{i=1}^{n_p} w_{I_p(i)}}, \qquad (2)$$

he vector of neighbor positions for position p, value of property j for the residue at position

position in the amino acid sequence. It is assumed that the geometrical neighborhoods are conserved, i.e. the neighbor positions and their distances for each sequence in the database are the same as those of the template. This assumption could be lifted in the future through the use of molecular dynamics refinement algorithms for the template structural information.

The second kernel, denoted as SeqNB, is the sequential kernel. This kernel is also described by Equations (1) and (2), but the number of neighbors around the residue, designated by n, is specified *a priori* along the sequence. A fixed average distance $\Delta = 1.3$ Å between any two consecutive residues is assumed and used to compute the weights $w_{I_p(i)}$. The distance between two residues separated by i positions in the sequence is $i\Delta$. For symmetric neighborhoods with $n/2$ neighbors on each side, the threshold T for the weight computation is $n\Delta/2$.

Note that Equation (2) explicitly defines the feature space and that the kernel in Equation (1) is computed as the inner product of these features. As a consequence, the Mercer condition (Vapnik, 1998) is satisfied and these kernels are valid kernels.

This classification problem is run on a previously developed computer program, ActiveSVM (Yu and Zavaljevski, 2003), which employs an efficient implementation of the active set method for solving the quadratic optimization, along with two regularization parameters to provide control for the sensitivity and specificity of the classifier (Veropoulos *et al.*, 1999).

3 RESULTS

3.1 Classification performance

The ActiveSVM algorithm with three different kernels was applied to four subgroups of immunoglobulin light chains. The geometric kernel is denoted by GeoNB(id), where id represents the PDB identification of the selected template. The sequential kernel is denoted by SeqNB(n), where n represents the number of sequential neighbors in the sequence segment of length $n+1$, with $n/2$ neighbors on each side. The third kernel in our implementation is LK. The LK is selected here to represent a standard kernel, as it was found to be the best kernel in our previous study (Zavaljevski *et al.*, 2002). Table 2 shows the performance based on the leave-one-out training/testing procedure. In addition to the overall classification error, Table 2 also presents the classifier sensitivity. For this application, sensitivity is considered more important than specificity. The cation performance

K	SeqNB(n)		GeoNB(id)			
	$n = 2$	$n = 4$	1BJM (λ_1)	1DCL (λ_2)	1LIL (λ_3)	1REI (κ_1)
3	**22**	22	39	29	29	31
8	**86**	82	68	71	75	71
4	35	**28**	39	39	41	39
3	63	**74**	58	63	53	63
5	43	37	35	33	**26**	31
5	45	55	55	55	**75**	55
3	30	26	30	36	30	34
2	69	75	69	64	67	64

ty.

in Table 2 show significant variability in kernel
or different subgroups. The best results for each
highlighted in bold face.

ging improves performance for the highly vari-
, it has a detrimental effect on the κ family.
us study, several critical point mutations were
κ family. When the sequences that have low
are averaged, averaging reduces information
he contrary, for sequences with high variabil-
can improve the signal to noise ratio and thus
sification. This is the case for the λ family, where
nsistently provides better performance than the

urprising result is the critical dependence of
nce of the geometric kernel on the selection
ural templates. A significant improvement is
only the λ_3 subgroup. However, it is prob-
e specific structural templates could improve
r the other groups as well. Without a struc-
, the classification error for the λ_3 subgroup
the structural template 1LIL reduces the error
a significant increase in sensitivity from 45 to
the best kernel for the λ_3 subgroup. The per-
lts using the structural templates from the other
lin light chains (1BJM, 1DCL and 1REI) are
d for this subgroup, when compared with the
ed by the LK. The best kernels for subgroups λ_1
qNB(2) and SeqNB(4), respectively, although

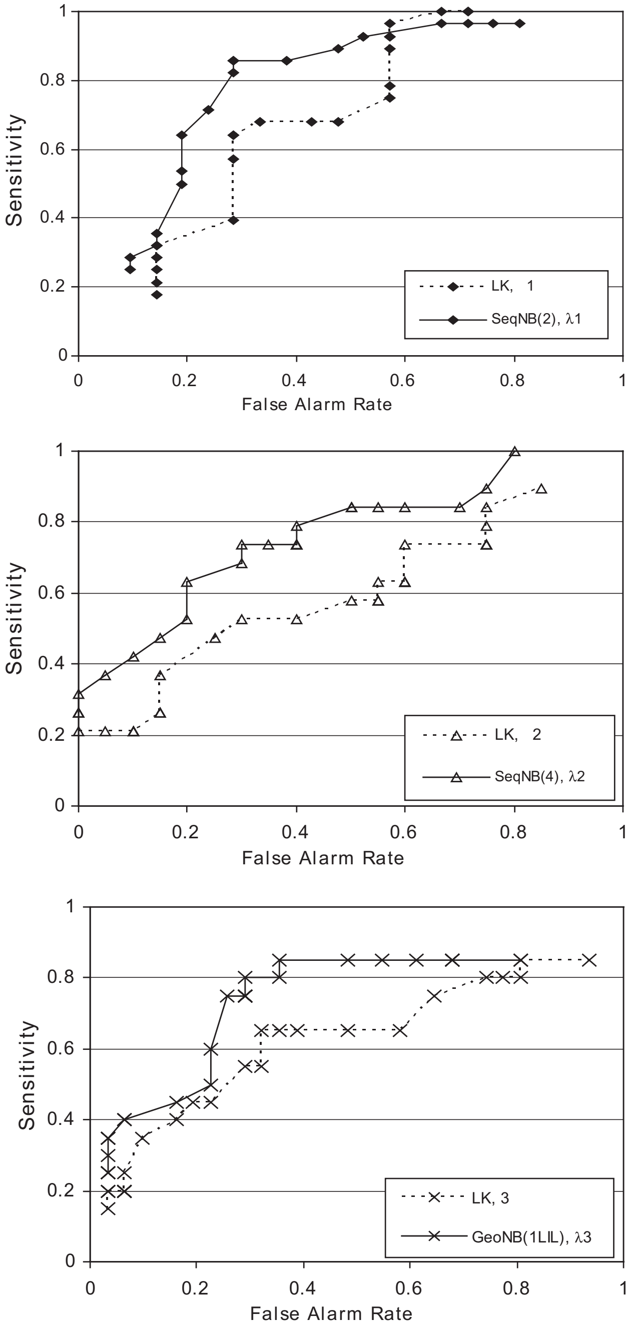

Fig. 1. ROC curve for the λ family.

nsitivity and significance results of the SVM classification

Error (%) Mean LK	SeqNB	Significance (P-value)[a]	Sensitivity (%) Mean LK	SeqNB	Significance (P-value)
37.9	28.3	6.0×10^{-11}	58.1	67.3	1.0×10^{-8}
37.3	33.1	1.0×10^{-4}	59.8	66.5	9.2×10^{-6}
43.3	38.0	4.5×10^{-3}	53.4	62.5	1.5×10^{-5}
40.2	38.2	5.4×10^{-2}	57.9	63.1	1.5×10^{-4}

tes the probability that the differences between two results are due to chance.

ta are pooled together to produce a dataset of 45 pathogenic proteins. Averaging is performed eighbors for the λ_1 and λ_2 subgroups and $n = 6$ he λ_3 subgroup, as Table 2 and additional sim-hown here) suggest a larger neighborhood for ge results over 50 such resamplings are given e Wilcoxon signed rank test (Myers and Well, med on the error and sensitivity results for each results show statistically significant improve-mance when sequential averaging is used in the mprovement in sensitivity is more significant. nce for the pooled data is worse than the per-he individual subgroups and is driven by the ation error of the λ_3 data.

e biological interpretations

results suggest that the mechanisms of amyloid ght be different for the λ_3 subgroup, perhaps fference in intrinsic propensity towards fibril

ght into possible mechanisms for this sub-d by calculating the scores χ_p^2 for position p

$$\left[\frac{(m^+_{pbj} - P_{pbj}m^+)^2}{P_{pbj}m^+} + \frac{(m^-_{pbj}P_{pbj}m^-)^2}{P_{pbj}m^-} \right], \tag{3}$$

ndex for the residue properties, B is the number partition the probability distribution for each he bin index, m^+_{pbj} is the number of pathogenic sition p with property j in bin b, m^-_{pbj} is the

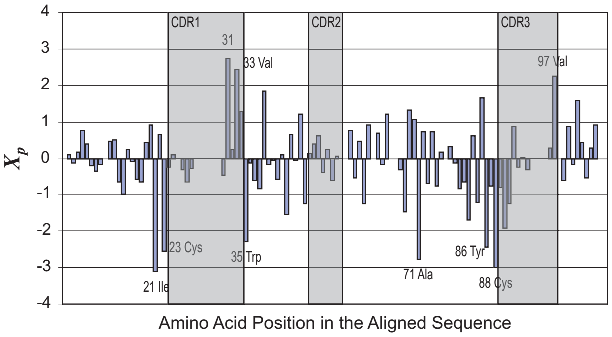

Fig. 2. Difference in χ_p^2 scores for each position without and with the 1LIL template.

where $\chi_p^2|_W$ denotes the score computed without the structural template and $\chi_p^2|_T$ denotes the score computed with the template 1LIL. This difference for the 120 amino acid positions is presented in Figure 2. The three highlighted regions are highly variable regions outside of the protein hydrophobic core, known as the complementarity-determining regions (CDRs). It has been suggested that amyloidosis is related to the protein hydrophobic core (Hoshino *et al.*, 2002). As a consequence, CDRs contribute less to amyloid formation. When the structural template is introduced, a significant increase in importance (denoted by a negative value in $\boldsymbol{X}_p$) of some positions outside of the CDRs can be observed. The importance of the variable regions is either suppressed or insignificant, except for a few positions in CDR3. The overall effect of the structural template improves amyloid discrimination, since the importance of regions that are expected to contribute to amyloid formation, such as hydrophobic regions, is

amyloidosis. The difference at the position of
as structural importance. This amino acid is
he classic 'tyrosine corner' in which it forms
ogen bond to the backbone carbonyl of Asp82.
e between Asp82 and Arg61 was implicated as
in κ family amyloidogenesis (Stevens, 2000).

USIONS

esults presented in this study indicate that modi-
he standard SVM kernels improve discrimin-
gn and pathogenic sequences in the presence
ence variability. Proper neighborhood struc-
ed for averaging of physicochemical properties
equence data. Thus, the major contribution of
the provision of an encoding strategy, which
special kernel functions tailored for this applic-
a mechanism for differential weighting of each
sequence that considers the interactions with
esidues. In this way, the encoding of each residue
e considers not only the amino acid type in that
so the location of the amino acid in the sequence.
cific case of immunoglobulin light chains, the
neighborhood structures among light chain
ght suggest various mechanisms of amyloid
each subgroup. For example, for the κ_1 sub-
sity for amyloid formation could be traced to
mutations at specific positions. For the λ_1 and
, short motifs of 3–5 amino acids in protein
ld indicate propensity for amyloid formation.
of mechanisms for the λ_3 subgroup is more
might suggest effects of non-local interactions
ormation, since a significant improvement is
this subgroup only when structural neighbor-
ded. However, due to very limited data, these
re only tentative and should be validated as
ental data become available. The importance
tabase of human immunoglobulin light chains
critical for determination of risk factors in the
e point mutations or sequence motifs. For lar-
more sophisticated methods for sequence motif
ld be implemented.
ection, i.e. the identification of the key amino
uence that are important in the characterization
to classification. In this manner, we could reduce the dimensionality of the encoding vector input to the SVM, reducing noise and potentially improving classification accuracy.

Another future direction for potentially improving protein classification is the computation of optimized structural templates. Strategies to be evaluated could include: creating models that incorporate all (human and non-human) sequences in the database and employing molecular dynamics for protein structure refinement. A second strategy addresses missing templates, i.e. germline representatives for which no structural representative currently exists in the database. In this case, models would be constructed by amino acid replacements of the most similar representative in the database, followed by energy minimization/molecular dynamics.

Many functionally diverse proteins share very similar folds. The distinction between amyloidogenic and non-amyloidogenic proteins is analogous to the distinction of proteins that have known function from those that do not have that function. Increasingly, due to increases in the number of known structures and improvements in recognition of fold at low levels of sequence similarity, it is possible to identify a probable fold. We anticipate that optimized incorporation of structural information with SVM algorithms could contribute significantly to the generation of functional hypotheses for proteins of currently unrecognized function.

ACKNOWLEDGEMENTS

The authors wish to express their gratitude to the anonymous referees for the useful comments and suggestions as well as interesting ideas for future work. The work presented in this paper was supported by the Laboratory Directed Research and Development Program of Argonne National Laboratory and by NIH grants DK43957 and AG18001 (FJS). Argonne National Laboratory is operated by the University of Chicago for the US Department of Energy. J.R. was supported by the Research Area Directorates of the US Army Medical Research and Materiel Command, Ft. Detrick, Maryland.

REFERENCES

Cai,C.Z., Han,L.Y., Ji,Z.L., Chen,X. and Chen,Y.Z. (2003) SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res.*, **31**, 3692–3697.

,Z. (2001) A novel method of protein secondary iction with segment overlap measure: support vector roach. *J. Mol. Biol.*, **308**, 397–407.

atou,H., Hagihara,Y., Hasegawa,K., Naiki,H. and 2) Mapping the core of the β_2-microglobulin amyloid exchange. *Nat. Struct. Biol.,* **9**, 332–336.

khans,M. and Haussler,D. (2000) A discriminative r detecting remote protein homologies. *J. Comput.* 114.

n,E., Weston,J. and Noble,W. (2002) Mismatch s for discriminative protein classification. *Neural Processing Systems 2002*, Vancouver, December

neider,G., Nehrens,D. and Wrede,P. (1994) A neural el for the prediction of membrane-spanning amino es. *Protein Sci.*, **3**, 1597–1601.

,M. and Serrano,L. (2004) Sequence determinants ibril formation. *Proc. Natl Acad. Sci. USA* **101**,

Well,A. (2003) *Research Design and Statistical* A, Mahwah, NJ.

04) Support vector machines applications in com- ology. In Schoelkopf,B., Tsuda,K. and Vert,J.-P. *l Methods in Computational Biology*. MIT Press, MA, pp. 71–92.

kman,L.J., Szpunar,M., Wunschl,C., Pokkuluri,P.R., tins,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 Sci.,* **8**, 509–517.

Sobolev,V., Sorokine,A., Priulsky,J., Abola,E.E. and Edelman,M. (1999) Automated analysis of interatomic contacts in proteins. *Bioinformatic*s, **15**, 327–332.

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

Stevens,F.J., Weiss,D.T. and Solomon,A. (1998) Structural base of light chain-related pathology. In Zanetti,M. and Capra,J.D. (eds), *The Antibodies*, Vol. 5. Harwood Academic Publishers, Australia, pp. 175–208.

Vapnik,V. (1998) *Statistical Learning Theory*. Wiley, New York.

Veropoulos,K., Cristianini,N. and Campbell,C. (1999) Controlling the sensitivity of support sector machines. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI99)*, Stockholm, Sweden.

Williams,S.C., Frippiat,J.-P., Tomlinson,I.M, Ignatovic,O., Lefranc,M.-P. and Winter,G. (1996) Sequence and evolution of the human germline V$_\lambda$ repertoire. *J. Mol. Biol.*, **264**, 220–232.

Yu,C. and Zavaljevski,N. (2003) *ActiveSVM User's Manual*, Argonne National Laboratory. Argonne, IL.

Zavaljevski,N., Stevens,F.J. and Reifman,J. (2002) Support vector machines with selective kernel scaling for protein classification and identification of key amino acid positions. *Bioinformatics*, **18**, 689–696.