JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 7, Number 6, 2000
Mary Ann Liebert, Inc.
Pp. 849–862

Representing and Reasoning about Protein Families Using Generative and Discriminative Methods

I.S. MIAN[1] and I. DUBCHAK[2]

ABSTRACT

This work addresses the issues of data representation and incorporation of domain knowledge into the design of learning systems for reasoning about protein families. Given the limited expressive capacity of a particular method, a mixture of protein annotation and fold recognition experts, each implementing a different underlying representation, should provide a robust method for assigning sequences to families. These ideas are illustrated using two data-driven learning methods that make use of different prior information and employ independent, yet complementary, projections of a family: hidden Markov models (HMMs) based on a multiple sequence alignment and neural networks (NNs) based on global sequence descriptors of proteins. Examination of seven protein families indicates that combining a generative (HMM) and a discriminative (NN) method is better than either method on its own. Biologically, human 4-hydroxyphenylpyruvic acid dioxygenase, involved in tyrosinemia type 3, is predicted to be structurally and functionally related to the glyoxalase I family.

1. INTRODUCTION

TECHNIQUES FOR MODELING PROTEIN FAMILIES are important tools for annotation and fold recognition studies. A family is defined usually on the basis of shared sequence and/or structural features and analysis of a multiple sequence alignment of its members. Using a set of sequences to discover features and derive models usually follows a three-step procedure. The first is hypothesis space selection: a model is chosen such that it characterizes the family. The second step is score function creation: a score function is designed such that, given a model and a set of positive and, if available, negative examples, a score of the examples (model) with respect to the model (examples) is returned. The third step is parameter estimation: an algorithm is developed which returns a good (preferably the best) model, given a set of examples. Increasing the true positive rate and minimizing the false positive and false negative rates of strategies for assigning sequences to families will require advances in two areas. The first is techniques for representing families (the inference method or inductive principal). The second is the accompanying algorithms used to reason about families (the learning methods). This work addresses the former issue of representation and reasoning rather than formal models and parameter estimation methods.

[1]Department of Molecular and Cell Biology (MS 74-197), Radiation Biology and Environmental Toxicology Group, Life Sciences Division, Lawrence Berkeley National Laboratory, Cyclotron Road, Berkeley, CA 94720.
[2]National Energy Research Scientific Computing Center (MS 84-171), Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720.

Two general types of models have been employed to study protein families: generative models such as hidden Markov models (HMMs), transformational grammars, and (object oriented and/or dynamic) Bayesian networks; and discriminative models such as neural networks (NNs), Support Vector Machines (SVMs) and decision trees (for recent reviews see Durbin *et al.* [1998], Baldi and Brunak [1998], Salzberg *et al.* [1998], Lander and Waterman [1998]). HMMs can handle sequences of variable length and, given a new sequence, return a (probabilistic) measure of how well the sequence can be assigned to the family under consideration. The use of discriminative methods has been hampered largely by the need to transform sequences of variable length into fixed-length input vectors representing features extracted from the original data. Thus, differences in the input–output requirements for each model type have spurred the more widespread use of generative methods in protein family modeling.

Differences in the expressive capacity of particular model types have lead to models being combined. This approach has resulted in significant improvements in pattern recognition tasks such as those encountered in computational biology. Hybrids of HMMs and NNs have been successful in a number of areas (see Riis and Krogh [1997], Baldi and Chauvin [1996], and references therein). In such cases, NNs are used primarily to estimate parameters of the HMM. In recent work (Jaakkola and Haussler, 1999; Jaakkola *et al.*, 1999), a generative model of a protein family (an HMM) served as the basis for deriving a kernel function for use in a discriminative method (an SVM). Such combination of model types yielded an improved method for classification of DNA and protein sequences. However, since the features used for discrimination were extracted from the generative model, the basic information employed by the two model types was derived from the same underlying representation of the family. Under such circumstances, any limitations in the HMM-based alignment would be propagated to the SVM. Thus, enhancing the performance of annotation and fold recognition systems will necessitate not only integrating generative and discriminative methods, but also employing different representations of the family.

This work demonstrates the feasability and utility of combining a generative model based on a representation in alignment space with a discriminative model based on a projection in single sequence space. In experiments designed to be illustrative rather than comprehensive, this mixture of experts is applied to seven selected families. An HMM (Durbin *et al.*, 1998; Baldi and Brunak, 1998; Salzberg *et al.*, 1998) for a family is trained using an iterative procedure that starts from a single representative sequence. The subsequences of family members that align to the HMM are used as input for an existing protein fold recognition system based on NNs and global sequence descriptors of sequences (Dubchak *et al.*, 1995; Dubchak *et al.*, 1997; Dubchak *et al.*, 1998; Dubchak *et al.*, 1999). Fold assignments for each member are combined to yield a final prediction for the family as a whole. Employing the HMM to pinpoint the subsequence most likely to be a member of the family improves the accuracy of fold assignment compared to that made using the full-length protein. For sequences in a database whose score against the generative model are below the threshold considered to indicate membership of the family, the application the discriminative model provides a novel mechanism for deciding whether the sequence could be a new, putative remote homolog.

In these proof-of-principal studies, the cyclin, glyoxalase I (Glo1), and glyoxalase II (Glo2) families have members with known structures. The other families are orotidine 5′-phosphate decarboxylase (DCOP), FHA, Band 4.1, and ATPases associated with various cellular activities (AAA) families. Biologically, the results indicate new members of each family and predictions of specific and more general protein folds for each family. For example, HMM-based analysis of *Escherichia coli* DCOP suggests proteins with a TIM barrel structure as being members of the family; subsequent NN-based fold recognition supports the proposition that DCOP has a TIM barrel fold. The human enzyme 4-hydroxyphenylpyruvic acid dioxygenase (HPPD) is predicted to be structurally related to the Glo1 family. This result provides insights into human tyrosinemia type 3, which is caused by a genetic deficiency of HPPD in tyrosine catabolism and is characterized by convulsion, ataxia, and mental retardation.

2. METHODS

This section summarizes the generative and discriminative methods utilized in this work and the details of how these existing approaches were applied to the families selected for study.

2.1. Protein family modelling using a generative method

Profile-based HMMs are statistical models of families that recast a multiple sequence alignment in a form suitable for searching sequence databases to identify remote homologs and for generating a multiple sequence alignment of a family (reviewed in Durbin *et al.*, 1998; Baldi and Brunak, 1998). HMMs have proved adept at increasing the number of proteins for which structural similarity can be inferred or implied from sequence similarity (for a discussion, see Eddy [1998]). They have two essential features: parameters for every position in the alignment which express the amino acid distributions and the insertion and deletion probabilities and a scoring function for sequences with respect to the model. The goal of training an HMM for a family, or training set, is to estimate the parameters of a model that assigns large likelihood to each sequence in the training set. HMMs have begun to emerge as an important method for large-scale sequence analysis (see, for example, Bateman *et al.* [1999], Karplus *et al.* [1998], Yu *et al.* [1998]).

In previous work (Mian, 1998; Mian, 1997; Dalgaard *et al.*, 1997), an initial sequence of interest and the PSI-BLAST (Altschul *et al.*, 1997) program were employed to identify a set of homologs of the query. Using the SAM implementation of HMMs (Hughey and Krogh, 1995), these sequences were utilized to train an initial HMM. This HMM was employed to identify remote homologs by iterative searching: statistically significant sequences found in one round of HMM searching were added to the training set and the expanded set used to retrain the HMM for the next round of searching. This procedure was repeated until no additional sequences were identified. New members of a family were identified by performing a database search: calculating the log-odds score (Barrett *et al.*, 1997; Eddy *et al.*, 1995) for every database sequence. The SAM NLL-NULL score measures how much more likely the HMM was to have generated the sequence than some competing model representing the universe of sequences as a whole rather than the sequence of interest. A specified level of significance of the score was used to assign new sequences to the family of interest (expected number of false positives = 0.01). HMM-based predictions of biological function borne out by subsequent experimental work include proteins predicted (Mian, 1997; Moster *et al.*, 1997) and shown to encode nucleases (Briggs *et al.*, 1998; Mitchell *et al.*, 1997; Zhang *et al.*, 1998; Huang *et al.*, 1998).

A multiple sequence alignment of a family indicates patterns of conservation and regions better able to tolerate insertions and deletions. Although HMMs capture local features, they cannot handle efficiently long-range correlations or overall features of the training set. Hence, additional approaches will be required to predict, for example, the folding class of a protein that has little sequence similarity to known proteins. The next section describes a method for fold recognition that addresses the inverse folding problem by considering global properties of a sequence rather than threading approaches that utilize pairwise potentials (Smith *et al.*, 1997).

2.2. Protein fold recognition using a discriminative method

Methods that use amino acid composition as the basis for assignment of a sequence to one of five broad structural classes can achieve a prediction accuracy above 70% (Chou and Zhang, 1995; Dubchak *et al.*, 1993). However, such methods perform less well as the number of classes and the similarity between classes increases. This difficulty arises because the parameter vectors of proteins from different classes become located closer in parameter space (Chou, 1989). Thus, prediction schemes based upon databases that classify proteins into larger numbers of folding classes require more complicated and detailed parameter sets.

A protein fold assignment technique has been developed that uses classifications containing from 20 to about 100 protein folds and NNs trained on vectors based on the physical and structural properties of the constituent amino acids (Dubchak *et al.*, 1995; Dubchak *et al.*, 1997; Dubchak *et al.*, 1998; Dubchak *et al.*, 1999). A protein is transformed from a string drawn from a 20-letter amino acid alphabet to one of the same length drawn from a different 3-letter alphabet. The six attribute alphabets currently used are predicted secondary structure, predicted solvent accessibility, polarity, polarizability, normalized van der Waals volume, and hydrophobicity. For a transformed sequence string, a 21-component vector capturing the overall property of the string is calculated as follows: the global composition of each letter in the alphabet (3 numbers), the frequencies with which the properties changed along the entire length of the string (3 numbers), and the distribution pattern of the property along the string (15 numbers). The seven

sets of parameters describing a single sequence include the six aforementioned 21-component vectors and a 20-component vector of amino acid composition. Given a fold classification scheme, seven NNs (one for each parameter set) are trained to distinguish one fold from the other folds.

Earlier work indicated that a consensus prediction was more reliable than individual predictions and that increasing the number of descriptors and accordingly trained NNs involved in the voting scheme improved prediction performance (Dubchak *et al.*, 1995). The fold assignment method was tested extensively both by cross-validation, where test sequences were not included in the training set, and on independent test sets (Dubchak *et al.*, 1999). The accuracy of a random correct prediction to a particular fold equaled $N/607$ and varied from $2/607 = 0.003(0.3\%)$ to $30/607 = 0.05(5\%)$ for the least and the most populated folds in the database. This performance is similar to earlier work where predictions were made in the context of assignment to 4–5 structural classes, whereas this method used a finer-grained classification consisting of 128 classes. In general, prediction performance varied widely for different folds. This observation highlights the importance of the quality and quantity of the training set (the number of representatives for a given fold).

To predict the fold of a new protein, the query sequence is transformed into the seven sets of aforementioned parameters and each set tested against the trained NNs. Predictions based on the individual attributes are used in a voting scheme for the final assignment of the protein to a fold (Fig. 1). Ideally, a protein sequence would be assigned to one fold and not assigned to all other folds in the library. In practice, however, more than one fold may be assigned if several folds are similar. No fold may be assigned if the number of votes required for prediction is too high or if the fold is not present in the library.

2.3. *Protein families studied*

The sequences of known or unknown structure chosen as representative members of the families are *Homo sapiens* cyclin G2A, essential for the control of cell cycle at the G2/M transition (SwissProt identifier CG2A_HUMAN; PDB identifier 1FIN; cyclin family); *Escherichia coli* orotidine 5′-phosphate decarboxy-

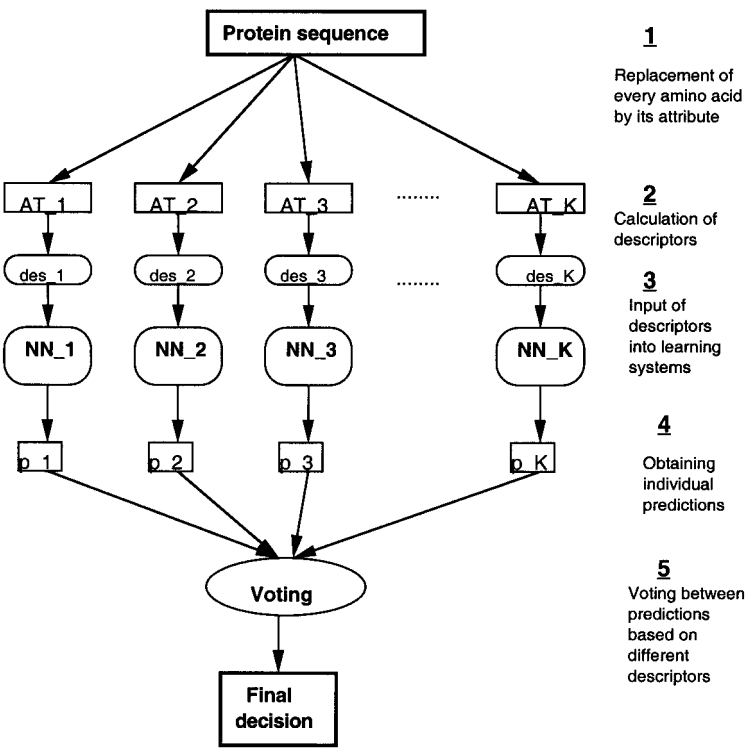

FIG. 1. The general NN-based fold assignment scheme. $AT_1, \ldots, AT_K$ are the amino acid attributes which replace the amino acid sequence. $des_1, \ldots, des_K$ are the vectors of descriptors based on the particular attribute. $p_1, \ldots, p_K$ are the predictions for each of the descriptors $des_1, \ldots, des_K$.

lase (DCOP_ECOLI; DCOP family); *H. sapiens* glyoxalase I (LGUL_HUMAN; 1FRO; Glo1 family); *Bacillus cereus* β-lactamase type II (BLA2_BACCE; 1BME, 1BMC; Glo2 family); *Mus mus* kinesin-like protein KIF1A (KF1A_MOUSE; FHA family); and *H. sapiens* focal adhesion kinase 1 (FAK1_HUMAN; Band 4.1 family).

2.4. HMM training

Each representative sequence was employed to train an HMM using the procedure summarized in Section 2.1. The studies used version 2.1.1 of SAM and the SWISS-PROT protein database release 36 and updates until August 1998. The August 1998 collection of AAA domain sequences, yeamob.pci.chemi.uni-tuebingen.de/AAA/Description.html, was used as is, i.e., no HMM was trained for the AAA family.

2.5. NN-based protein fold recognition

For the cyclin, DCOP, Glo1 and Glo2 families, two separate fold assignments were made for each member. Alignment of a sequence to an HMM indicates the subsequence most likely to represent the domain modeled by the HMM. However, the sequences for most family members are much longer. Thus, assignments were made for the full-length protein as well as for the aligned subsequence. For the FHA, Band 4.1, and AAA families, predictions were made for only the subsequences. Every sequence of interest was transformed into the seven parameter vectors described earlier.

Predictions were made using a previously trained system that recognizes protein folds in the context of the comprehensive Structural Classification of Proteins (SCOP) (Murzin *et al.*, 1995). For NN training, a SCOP subset based on the 35% cutoff PDB select set (Hobohm and Sander, 1994) was built. The resulting database contains 607 nonhomologous proteins representing 128 folds of SCOP (Dubchak *et al.*, 1999). Seven NNs were trained to distinguish one fold from the other 127 folds; the total number of trained NNs is 128×7. In Fig. 1 therefore, $K = 7$ and the entire scheme is repeated 128 times, once for each of the 128 folds in the SCOP subset used. To make a prediction, each vector was tested against the trained NNs and assigned to one of 128 folds depending upon the decisions of the seven NNs for each fold. A sequence is predicted to have a particular fold only if more than half of the predictions are positive. The number of votes used for decision making correlates with the selectivity and sensitivity of the prediction, i.e., the accuracy of the prediction and the probability of avoiding a correct prediction. Tests showed that reliable prediction could be achieved only if at least five out of seven possible votes give positive predictions. Although four positive votes present a plausible hypothesis, such predictions are considered to be of intermediate reliability.

Every protein sequence tested gets 128 individual positive or negative assignments to all 128 folds. Predicting several folds as well as predicting none can take place. Prediction can be made if only one fold is positively predicted or if the number of votes for one fold is higher than for others. If none of the 128 folds satisfies these conditions, the fold is listed as "Unknown," i.e., any prediction in the context of the SCOP subset is of insufficient confidence. Different folds of the training SCOP subset have different numbers of members; the reliability of prediction is higher for folds with large numbers of members because NN training results in better generalization (Hertz *et al.*, 1991). The most reliable assignments are those when the test sequence is recognized as belonging to one of the folds containing six or more proteins in the training set. There were 37 such folds in the training set.

3. RESULTS

3.1. HMM based detection of new family members

For the seven families studied, HMMs were trained and used to identify new members. Each sequence assigned to a family was used as the input for the NN-based fold prediction method. Table 1 gives the HMM NLL-NULL score and predicted fold for the aligned subsequence for each cyclin family member (similar tables for the other families are available from the authors). Although a number of sequences not previously described as belonging to the families were identified, only one example is described because of its biomedical importance.

Table 1. HMM and Fold Prediction Studies of the Cyclin Family[a]

Sequence	Predicted fold	Length	NLL-NULL
SceClb2a	β OB-fold	98	−106.290
SceClb1a	unknown	98	−103.250
HsaG2Aa	unknown	98	−103.240
SpoG23a	unknown	98	−103.170
SceClb6a	α EF-hand	98	−102.930
SpoCIG1a	α 4-helical cytokines	98	−96.980
SceClb5a	α/β Thioredoxin-like	98	−96.190
XleG2B2a	β Immunoglobulin-like	98	−96.150
SceClb3a	unknown	97	−94.590
CviG2Ba	α 4-helical cytokines	98	−92.920
CelG2Ba	unknown	98	−91.280
SpoCIG2a	unknown	98	−91.260
SceClb4a	α EF-hand	97	−89.770
HsaG2B1a	α/β Thioredoxin-like	98	−86.570
CelG2A1a	α/β Thioredoxin-like	98	−84.060
SpoPUC1	unknown	100	−81.820
CelCyc3	β Lipocalins	99	−80.650
AfuTF2Ba	unknown	95	−79.840
PfuTF2Bb	unknown	95	−78.640
PhoTF2Ba	unknown	95	−77.610
HsaG1SD3	unknown	98	−77.230
CelG2B3a	unknown	99	−77.060
SshTF2Ba	β Immunoglobulin-like	95	−75.840
Hh8vCyc	β-Trefoil	98	−74.800
PfuTF2Ba	β Immunoglobulin-like	95	−74.700
AfuTF2Bb	unknown	94	−74.600
ScePcl1	α/β Ribonuclease H-like motif	101	−72.600
HsaT1a	unknown	112	−71.400
SceCcl1a	β Immunoglobulin-like	95	−70.700
SshTF2Bb	unknown	94	−69.260
SceClb2b	unknown	95	−68.830
HsaG2G1	β OB-fold	100	−67.490
ScePcl2	unknown	102	−66.090
SceClb1b	α/β Thioredoxin-like	95	−64.930
AthTF2Ba	α Globin-like	95	−64.400
HsaTF2Ba	unknown	94	−64.250
SceCln2	unknown	123	−64.180
SpoG23b	β Lipocalins	94	−64.100
HsaCycHa	unknown	104	−63.890
SpoCIG2b	unknown	95	−63.230
PmaPREG	unknown	99	−63.170
GciG1S	unknown	103	−63.100
CelG1ACa	unknown	106	−63.060
HsaG2B1b	unknown	94	−63.050
SceCln1	unknown	123	−62.490
HsaG2F	NAD(P)-binding Rossmann	98	−62.170
PhoTF2Bb	α DNA-binding 3-helical bundle	94	−61.480
AthD2	α/β Thioredoxin-like	101	−61.420
SceCln3	unknown	100	−61.270
XleCycHa	β OB-fold	93	−60.620
Mh6vCyc	α/β NAD(P)-binding Rossmann	97	−60.040
SceClb6b	β OB-fold	95	−59.600
CviG2Bb	unknown	95	−59.220
ScePc19	unknown	101	−58.670

(*continued*)

TABLE 1. (CONTINUED)

Sequence	Predicted fold	Length	NLL-NULL
SceClb3b	unknown	94	−58.640
CelTF2Ba	unknown	95	−58.160
XleG2B2b	unknown	94	−57.840
SceSRB11a	β OB-fold	103	−57.730
KluTF2Bb	β Viral coat and capsid pro	108	−57.530
NcrPREG	unknown	99	−57.180
MmuG1SCa	unknown	112	−56.790
HsaTF2Bb	unknown	95	−54.560
CelCyc2	α/β Thioredoxin-like	110	−53.850
ScePcl7	β Lipocalins	110	−53.230
SceClb4b	unknown	94	−53.110
AthCycC1	β Lipocalins	99	−52.910
ScePcl5	unknown	103	−52.800
SpoMCS2a	β Immunoglobulin-like	96	−52.670
KluTF2Ba	unknown	94	−52.250
SceClb5b	unknown	93	−51.800
SpoCIG1b	α/β Thioredoxin-like	94	−50.490
SceTFIIIBb	unknown	98	−50.480
ScePcl10	unknown	106	−49.990
SpoC20F10	β Immunoglobulin-like	100	−49.980
CelTF2Bb	α/β Thioredoxin-like	97	−49.570
We2Cyca	unknown	100	−49.300
SpoC19E9	unknown	101	−49.140
ScePho80	β Immunoglobulin-like	99	−48.690
HsaRB1b	β OB-fold	104	−48.650
AthCYCb	unknown	95	−48.050
XleRBb	unknown	115	−47.920
SceTFIIBa	α/β Ribonuclease H-like motif	94	−47.770
CelCyc1	unknown	97	−47.470
SceCTK2a	α 4-helical cytokines	108	−46.610
CelYOJ1	unknown	98	−46.350
CelC52E4b	α/β Ribonuclease H-like motif	98	−45.970
SpoCyc	unknown	108	−45.960
GgaRBb	unknown	115	−45.470
HsaRBb	α 4-helical cytokines	115	−45.420
ScePcl8	β OB-fold	107	−45.420
ScePcl6	unknown	161	−45.360
CelYL34a	unknown	117	−45.180
CelYL33a	unknown	114	−45.160
SpoC2G2	unknown	97	−43.950
CelC52E4a	unknown	104	−43.910
SceCTK2b	β OB-fold	95	−43.820
SceClg1	α/β Ribonuclease H-like motif	104	−43.790
MmuG1SCb	β OB-fold	92	−43.770
AthTF2Bb	unknown	92	−43.350
SceTFIIIBa	unknown	95	−42.930
CelG2B3b	unknown	97	−42.830
HsaRB2b	unknown	104	−41.950
CelYL33b	β Lipocalins	105	−41.240
SceTFIIBb	unknown	108	−40.950
HsaCycHb	unknown	102	−40.120
We2Cycb	unknown	102	−39.830
SceSRB11b	β Immunoglobulin-like	119	−39.650

(*continued*)

TABLE 1. (CONTINUED)

Sequence	Predicted fold	Length	NLL-NULL
PbcTF2Ba	unknown	100	−39.350
CelZK678	α/β Thioredoxin-like	82	−38.050
XleCycHb	unknown	102	−38.050
DmeRBb	unknown	97	−36.960
CelF09G2	β Immunoglobulin-like	116	−36.930
WelCyc	unknown	100	−35.250
HsaRB1a	α 4-helical cytokines	124	−35.050
CelRBb	unknown	100	−33.500
CelG2A1b	α/β Thioredoxin-like	95	−32.250
GgaRBa	α/β (TIM)-barrel	133	−31.470
HsaT1b	unknown	101	−31.300
HsaG2Ab	unknown	96	−31.120
ZmaRB1a	α 4-helical cytokines	108	−30.350
HsaRBa	unknown	131	−29.350
CelG2Bb	α/β NAD(P)-binding Rossmann	108	−28.730
ZmaRB1b	unknown	128	−28.180
CelYL34b	β Lipocalins	106	−26.730
CelG1ACb	unknown	105	−25.860
DmeRBa	α/β P-loop	145	−24.900
SpoMCS2b	β Lipocalins	120	−24.550
PbcTF2Bb	unknown	96	−23.480
XleRBa	α Ferritin like	132	−23.050
HsaRB2a	unknown	123	−19.070
SceCcl1b	β OB-fold	135	−16.210

[a] For each family member, the sequence name, predicted fold for the aligned subsequence, full length of the protein and HMM NLL-NULL score is shown.

The Glo1 family includes members from two different Prosite entries (PDOC00078 and PDOC00720). The first entry encompasses the extradiol ring-cleavage dioxygenases which catalyse the incorporation of both atoms of molecular oxygen into substrates. Examples include catechol 2,3-dioxygenase or metapyrocatechase (SwissProt sequences NAHH_PSEPU, XYLE_PSEAE, DMPB_PSESP, PHEB_BACST, MPC2_ALCEU); 3–methylcatechol 2,3–dioxygenase (TODE_PSEPU); biphenyl–2,3–diol 1,2–dioxygenase (BPHC_BURCE); 1,2–dihydroxynaphthalen dioxygenase (NAHC_PSEPU) and 2,2′,3–trihydroxybiphenyl dioxygenase. The second Prosite entry covers glyoxalase I (LGUL_ECOLI, LGUL_HUMAN) which catalyse the transformation of methylglyoxal and glutathione into S-lactoylglutathione. The results suggest that 4-hydroxyphenylpyruvate dioxygenases (4HPPDs) are structurally and functionally related to other members of the Glo1 family. Figure 2 shows an alignment of this family including human 4HPPD (HPPD_HUMAN) which is involved in the catabolism of tyrosine. Defects in human 4HPPD are the cause of type III tyrosinemia (MIM [McKusick, 1997] 276710). This observation provides insights into the structure and mechanism of action of this enzyme and thus site-directed mutagenesis and other experiments aimed at probing the prediction.

3.2. Fold recognition

The structure of the family as a whole was characterized by tabulating the repertoire of predicted protein classes/folds for each family member. Table 2 shows SCOP class assignments for the cyclin, DCOP, Glo1 and Glo2 families for both the full length proteins and the aligned subsequences. Table 3 shows the SCOP fold assignments for the aligned subsequences of these families and the Band 4.1, FHA, and AAA families.

For the cyclin, Glo1, and Glo2 families, only the aligned subsequences but not the full length proteins are assigned to the "Unknown" group (Table 2). Since these folds were not part of the SCOP set used to

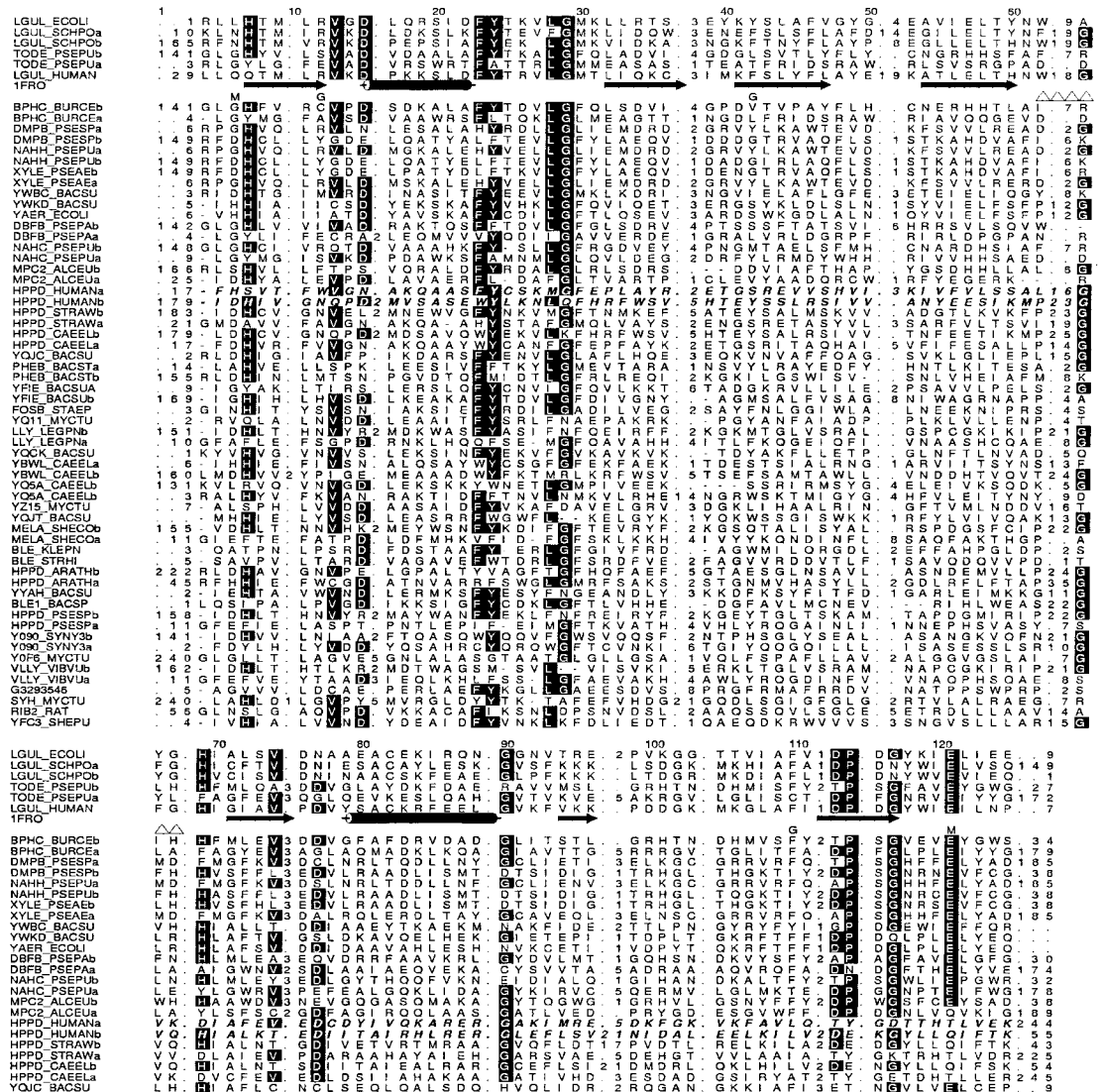

FIG. 2. An HMM-generated alignment of selected members of the Glo1 family. Some sequences contain two copies of the domain family shown (for example LGUL_SCHPOa and LGUL_SCHPOb). Residues conserved in the majority of sequences are highlighted. Arrows and cylinders indicate the locations of the β-strands and α-helices in the structure of human glyoxalase I (sequence LGUL_HUMAN; PDB code 1FRO) (Cameron *et al.*, 1997). "M" and "G" denote residues that interact with the metal ion or glutathione and Δ the interdomain region (the alignment shows two copies of the repeated $\beta\alpha\beta\beta\beta$ structural unit). The sequence of human 4-hydroxyphenylpyruvic acid dioxygenase (HPPD_HUMANa, HPPD_HUMANb), predicted to be a new member of the Glo1 family, is shown in italic. Other family members of known structure are not shown (BPHC_PSEPS, PDB 1HAN; BPHC_PSES1, PDB 1DHY).

TABLE 2. RESULTS OF FOLD ASSIGNMENTS FOR THE CYCLIN, DCOP, GLO1 AND GLO2 FAMILIES[a]

Family	SCOP Class	α	β	α/β	$\alpha+\beta$	Unknown	Total
cyclin	α	4	3	1	0	1	9
Complete: 271 (95, 786; 150)	β	0	6	2	0	14	22
Aligned subsequence: 102 (82, 161; 11)	α/β	4	8	8	0	21	41
SCOP class: α	$\alpha+\beta$	0	0	0	0	0	0
	Unknown	3	11	8	0	37	59
	Total	11	28	19	0	73	131
DCOP	α	0	0	1	0	0	1
Full length: 341 (198, 937; 236)	β	0	5	4	0	2	11
Aligned subsequence: 230 (156, 346; 39)	α/β	1	10	119	0	49	179
SCOP class: α/β	$\alpha+\beta$	0	0	0	0	0	0
	Unknown	5	20	80	0	55	160
	Total	6	35	204	0	106	351
Glo1	α	0	0	0	0	1	1
Full length: 284 (116, 732; 89)	β	0	12	1	0	13	26
Aligned subsequence: 125 (103, 271; 19)	α/β	6	35	45	0	94	180
SCOP class: $\alpha+\beta$	$\alpha+\beta$	0	0	0	0	0	0
	Unknown	1	13	8	0	53	75
	Total	7	60	54	0	161	282
Glo2	α	0	0	0	0	0	0
Full length: 380 (200, 1425; 185)	β	0	14	4	0	10	28
Aligned subsequence: 202 (154, 307; 32)	α/β	0	24	51	0	71	146
SCOP class: $\alpha+\beta$	$\alpha+\beta$	0	0	0	0	0	0
	Unknown	2	21	36	0	47	106
	Total	2	59	91	0	128	280

[a] Two predictions at the SCOP class level were made for each family member: the subsequence that aligns to the HMM (column totals) and the full length protein (row totals). As an illustration of how to read the table, consider the 131 cyclinfamily members. For 9 members, the aligned subsequences were assigned to the α class. However, for only 4 of these members was the full-length sequence assigned to α class (3 β, 1 α/β, 1 Unknown). Fifty-nine (73) aligned subsequences (full length protein) were assigned to the Unknown class. The column headed "Family" gives the mean length of the sequences together with the minimum length, maximum length, and standard deviation in parenthesis. As might be expected, the range of lengths of aligned subsequences is considerably smaller than the full-length proteins. The SCOP classes of families where one or more members have known three-dimensional structures are given. Since none of these were part of the original training set used to develop the NN-based fold assignment system, predictions for these families should fall into the "Unknown" class.

train the NN classifiers, these assignments are considered to be correct. Recall that the NNs were trained for recognition at the fine-grained fold rather than the coarse-grained class level given in Table 2. For the DCOP family, the aligned subsequences are assigned to the α/β class. Although the structure of *E. coli* DCOP is unknown, database searches with the DCOP HMM revealed family members having known TIM barrel structures (TRPC_ECOLI, PDB identifier 1PII; TRPC_SULSO, 1JUK; IMDH_TRIFP, 1AK5; TRPA_SALTY, 2TSY). Of the 351 sequences predicted to belong to the DCOP family (Table 2), 58.1% are predicted to have an α/β fold and 30.2% an Unknown fold (21.1% TIM-barrel and 25.9% NAD(P)-binding Rossmann). Thus, the HMM- and NN-based results are self-consistent leading to the prediction of a TIM-barrel fold for the DCOP family.

Excluding the Unknowns, 21.3% of the cyclin family members are predicted to be all β. Although this is likely to reflect deficiencies in the fold prediction technique employed, it might be indicative of an alternative structure for the cyclin family. There is precedence for such a possibility. Propagation of the infectious agent called a prion is thought to involve the conversion of the cellular protein PrPc from a predominantly α-helical to a β-sheet structure (reviewed in Tatzelt *et al.* [1998]). Whether a similar

TABLE 3. PREDICTED PROTEIN FOLDS FOR THE ALIGNED SUBSEQUENCES OF FAMILIES CHARACTERIZED BY HMMs[a]

	Fold	Glo1	Glo2	DCOP	cyclin	Band 4.1	FHA	AAA
α	Four-helical up-and-down	1	0	0	0	0	1	0
	Globin-like	6	1	6	1	0	0	0
	4-helical cytokines	0	0	0	6	0	0	1
	EF-hand	0	1	0	2	0	0	0
	DNA-binding 3-helical bundle	0	0	0	1	0	1	0
	Ferritin-like	0	0	0	1	3	0	0
	Percent of total	2.5	0.7	1.7	8.4	3.3	1.6	0.3
β	Acid proteases	2	0	0	0	0	0	0
	Cupredoxins	3	2	5	0	0	0	0
	Immunoglobulin-like	21	37	26	9	1	9	7
	Lipocalins	4	7	0	7	1	1	0
	Reductase/elongation factor	1	0	0	0	0	0	0
	SH3-like barrel	1	0	0	0	0	0	0
	Trypsin-like serine proteases	2	0	0	0	0	0	0
	Single β-helix	1	1	0	0	1	1	3
	ConA-like lectins/glucanases	0	2	0	0	0	0	0
	Viral coat and capsid protein	1	7	3	1	0	1	0
	β-Trefoil	0	1	0	1	1	1	0
	OB-fold	24	2	1	10	30	9	0
	Percent of total	21.3	21.1	10.0	21.3	37.4	18.2	3.7
α/β	(TIM)-barrel	10	42	75	1	12	0	67
	FAD-binding motif	3	5	6	0	0	2	3
	Flavodoxin-like	1	2	2	0	0	0	3
	Hydrolases	3	1	1	0	0	0	0
	NAD (P)-binding Rossmann	15	34	91	3	1	0	31
	P-loop	4	1	22	1	1	0	22
	Ribonuclease H-like motif	6	2	1	4	4	1	1
	Thioredoxin-like	12	3	0	10	0	0	1
	Periplasmic binding protein	0	1	6	0	0	1	0
	Percent of total	19.2	32.5	58.1	13.7	19.8	3.3	46.9
$\alpha + \beta$	β-grasp	0	0	0	0	0	1	0
	Metzincins	0	0	0	0	0	1	0
	Percent of total	0.0	0.0	0.0	0.0	0.0	1.6	0.0
Unknown		161	128	106	73	36	91	134
	Percent of total	57.1	45.7	30.2	55.7	39.6	75.2	49.1
Total number of family members		282	280	351	131	91	121	273

[a] For example, 7 Glo1 family were assigned to the α class (1 "Four-helical up-and-down", 6 "Globinlike") but this constitutes only 2.5% of the 282 members of the family.

conformational transition of cyclin family members occurs *in vivo* and under what specific conditions remains to be seen.

Of the FHA family, 75.2% of its members are predicted to have an Unknown fold and 18.2% a β fold. A "novel" fold is predicted for this family. Amongst the Band 4.1 family, 39.6% are predicted to have an Unknown fold and 37.4% a β fold. A "novel" β fold is predicted for this family. For the AAA family, 55.7% are predicted to have an Unknown fold and 21.3% a β fold. A "novel" β fold is predicted for this family.

3.3. Knowledge engineering of protein families

The two primary features of knowledge representation are the information that is made explicit and how the information is physically encoded for subsequent use. The aforementioned generative and discriminative approaches exhibit distinct differences in these areas. In the NN-based method, the original sequences are transformed by a preprocessing step to give a new set of variables encoding global aspects of sequences. These variables, calculated from positive and negative training examples, are treated as the input to the classifiers. The outputs, binary decisions, are postprocessed to yield the final decision. In contrast, sequences are used directly to estimate the HMM, a formal state sequence model that is a first-order Markov chain. The output of an HMM is a probability (log-odds score) since it describes a probability distribution over a (potentially) infinite number of sequences.

The precise form of the knowledge and prior knowledge plays a key role in generating "good" representations and thus the subsequent performance of the model under consideration. In the NN work, prior knowledge is incorporated via the precise form of the global sequence descriptor, the choice of amino acid properties used, and the classification of folds. The HMMs model the incomplete or inexact nature of the data by, for example, using methods for estimating probabilities of amino acids and transitions given small samples (Sjölander *et al.*, 1996). Prior knowledge also plays a role in which columns of a multiple sequence are assigned to the match states of the HMM. For both model types, family members that have similar representations are "closer" together than nonfamily members (as judged by the log-odds score of the HMM and the dissimilarity measure between global sequence descriptors by the NNs).

For the task of assigning a sequence to a folding class, representing a protein sequence by its global characteristics has several advantages. The information content of sequences is reduced to a small, more manageable number of numerical parameters that can be calculated quickly and easily from the primary sequence. Since these parameters can be employed efficiently by various machine-learning methods, predicting the folding class for a new sequence can be performed rapidly and automatically. However, an apparent disadvantage of all the knowledge-based prediction methods, including this one, is incompleteness in the database of folding classes. About fifty protein folds not observed before have appeared among over one thousand new structures during the last two years. Thus, the trained NNs used in this work represent an incomplete, description of protein fold space.

For the task of characterizing the sequence features of a family likely to be biologically important, for example putative active site residues, an HMM-generated alignment is more appropriate. Thus, the two techniques used here provide views of the families at complementary resolutions and permit different inferences to be made.

4. DISCUSSION

Although only a limited set of experiments were performed, the results of combining two specific generative and discriminative methods are encouraging. The flexibility, sensitivity, and specificity of HMMs in creating statistical models for families at the sequence level was exploited and combined with the generalization power of NNs in predicting protein folding class using global sequence descriptors. The results demonstrate the benefits of a) employing HMMs as a preprocessing mechanism for partioning a protein into the segments for which an NN-based fold prediction is to be made and b) performing a consensus fold prediction for a family as a whole that is based on individual predictions for each member. As demonstrated by the DCOP family, combining evidence from different models and model types enhances the ability to assign accurately sequences to a fold class. Such an approach ameliorates the inherent limitations in the expressive capability of the HMM and NN techniques used here.

Combining the HMMs and NNs in a protein annotation and fold prediction scheme has a number of advantages. A carefully collected and refined set of HMM families of known structure provides a good database for NN training. Furthermore, it allows development of new global descriptors based on the information contained in an HMM. An HMM database of families would include more distant sequence homologs thereby increasing fold prediction accuracy. Similarly, incorporation of structurally similar proteins as defined by the discriminative method could be used to improve the ability of HMMs to detect remote sequence homologs. This approach is modular and flexible: the fold classification scheme and library of families can be developed and applied separately. The whole scheme can be extended or modified

by adding any number of folds, new protein sequence descriptors, and machine-learning systems. Since the scheme has parallel as well as sequential operations, it is ideally suited for massively parallel processing. The whole process can include both semi- and fully automated steps. Increasing the number of descriptors leads to better, more accurate fold predictions without great added computational expense. Overall, this approach could lead to a system for protein annotation and fold recognition more accurate than either method on its own.

It should be noted that even an HMM which maximizes the likelihood of the training data is inherently an imperfect model for a family. This issue was illustrated in previous work on a family 1 glycosidase domain (Mian, 1998). That work highlighted a need to develop a set of HMMs, each designed to capture different aspects of the sequences and tailored to address different questions. An HMM was trained to capture the core elements of the $(\beta/\alpha)_8$ barrel global fold and the residues involved in recognizing the carbohydrate substrate. However, if the goal was to detect more remote homologs and/or characterize the $(\beta/\alpha)_8$ barrel, the connecting regions within and between the β/α repeats could be modeled implicity by converting them to insertions leaving only the β-strands and α-helices of the core barrel. Such a strategy would be most suitable for identifying distant relationships by merging specific glycosidase families and superfamilies in an effort to approximate an archetypal glycosidase fold. The conflicting demand to have a single, specific, and sensitive model represent a diverse set of family members is accompanied by an inability to develop a single model type of sufficient expressive capability. Hence, alternative models of a given type and different model types should be created for a single family. Thus, a "one-family–one-model" strategy should be revised to one of "one-family–many-models."

The representations employed in this work provide neither a comprehensive nor a complete description of all the features that are important in protein families. For example, subcellular localization, phosphorylation state, and so on are not modeled. Developing robust, semi-automated, large-scale, high-throughput protein annotation and fold recognition systems will require designing a mixture of experts containing a variety of learning methods implementing different representations. For example, representations such as Markov random fields (Stultz *et al*., 1993) would complement the alignment and global sequence description agents employed here. Recent work (M.L. Chow and I.S. Mian) suggests that SVMs could replace NNs in the current fold assignment scheme and, in conjunction with novel types of global sequence descriptor, increase the accuracy of the fold prediction scheme.

ACKNOWLEDGMENTS

This work was supported by the Director, Office of Energy Research, Office of Health and Environmental Research, Division of the US Department of Energy under Contract No. DE-AC03-76F00098. The data are available upon request.

REFERENCES

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. 1997. *Nucl. Acids Res.* 25, 3389–3402. The WWW-interface at the NCBI is available at URL www.ncbi.nlm.nih.gov/cgi-bin/BLAST/nph-psi_blast.

Baldi, P., and Brunak, S. 1998. *Bioinformatics: The Machine Learning Approach.* MIT Press, Cambridge, MA.

Baldi, P., and Chauvin, Y. 1996. *Neural Computation* 8, 1541–1565.

Barrett, C., Hughey, R., and Karplus, K. 1997. *CABIOS* 13, 191–199.

Bateman, A., Birney, E., Durbin, R., Eddy, S., Finn, R., and Sonnhammer, E. 1999. *Nucl. Acids Res.* 27, 260–262.

Briggs, M., Burkard, K., and Butler, J. 1998. *J. Biol. Chem.* 273, 13255–13263.

Cameron, A., Olin, B., Ridderström, M., Mannervik, B., and Jones, T. 1997. *EMBO J.* 16, 3386–3395.

Chou, K., and Zhang, C. 1995. *Critical Reviews in Biochemistry and Molecular Biology* 30, 275–349.

Chou, P. 1989. In *Prediction of Protein Structure and the Principles of Protein Conformation*, Plenum Press, New York.

Dalgaard, J., Moser, M., Hughey, R., and Mian, I. 1997. *J. Comp. Biol.* 4, 193–214.

Dubchak, I., Holbrook, S., and Kim, S.-H. 1993. *Proteins* 16, 79–91.

Dubchak, I., Muchnik, I., Holbrook, S., and Kim, S. 1995. *Proc. Nat. Acad. Sci.* 92, 8700–8704.

Dubchak, I., Muchnik, I., and Kim, S. 1997. *ISMB* 5, 104–107.
Dubchak, I., Muchnik, I., and Kim, S.-H. 1998. *Microbial and Comparative Genomics* 3, 171–175.
Dubchak, I., Muchnik, I., Mayor, C., Dralyuk, I., and Kim, S.-H. 1999. *Proteins* 35, 401–407.
Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. 1998. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press, Cambridge.
Eddy, S. 1998. *Bioinformatics*, 14.
Eddy, S., Mitchison, G., and Durbin, R. 1995. *J. Comp. Biol.* 2, 9–23.
Hertz, H., Krogh, A., and Palmer, R. 1991. *Introduction to the theory of neural computation.* Addison-Wesley, Redwook City California.
Hobohm, U., and Sander, C. 1994. *Protein Science* 3, 522–524.
Huang, S., Li, B., Gray, M., Oshima, J., Mian, I., and Campisi, J. 1998. *Nature Genetics* 20, 114–116.
Hughey, R., and Krogh, A. (1995). Technical Report UCSC-CRL-95-7 University of California, Santa Cruz Computer and Information Sciences Dept., Santa Cruz, CA 95064.
Jaakkola, T., Diekans, M., and Haussler, D. Unpublished.
Jaakkola, T., and Haussler, D. Unpublished.
Karplus, K., Barrett, C., and Hughey, R. 1998. *Bioinformatics* 14, 846–856.
Lander, E., and Waterman, M., eds. 1998. *Calculating The Secrets of Life: A Mathematician's Introduction to Molecular Biology.* National Academy Press.
McKusick, V. The OMIM database is available at URL www3.ncbi.nlm.nih.gov/omim/.
Mian, I. 1997. *Nucleic Acids Research* 25, 3187–3195.
Mian, I. 1998. *Blood Cells, Molecules, & Disease* 24, 83–100.
Mitchell, P., Petfalski, E., Shevchenko, A., Mann, M., and Tollervey, D. 1997. *Cell* 91, 457–466.
Moser, M., Holley, W., Chatterjee, A., and Mian, I. 1997. *Nucleic Acids Research* 25, 5110–5118.
Murzin, A., Brenner, S., Hubbard, T., and Chothia, C. 1995. *J. Mol. Biol.* 247, 536–540.
Riis, S., and Krogh, A. 1997. In *Proceedings of ICASSP'97.* IEEE New York.
Salzberg, S., Searls, D., and Kasif, S., eds. 1998. *Computational Methods in Molecular Biology.* Elsevier Science.
Sjölander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I., and Haussler, D. 1996. *CABIOS* 12, 327–345. More up-to-date information is available from www.cse.ucsc.edu/research/compbio/dirichlets/index.html.
Smith, T., Lo Conte, L., Bienkowska, J., Gaitatzes, C., Rogers Jr., R., and Lathrop, R. 1997. *J. Comp. Bio.* 4, 217–225.
Stultz, C., White, J., and Smith, T. 1993. *Protein Science* 2, 305–314.
Tatzelt, J., Voellmy, R., and Welch, W. 1998. *Cellular and Molecular Neurobiology* 18, 721–729.
Yu, L., White, J., and Smith, T. 1998. *Protein Science* 7, 2499–2510.
Zhang, X., Zhu, L., and Deutscher, M. 1998. *J. Bact.* 180, 2779–2781.

Address correspondence to:
I.S. Mian
Department of Molecular and Cell
Biology (MS 74-197)
Lawrence Berkeley National Laboratory
1 Cyclotron Road
Berkeley, CA 94720

E-mail: SMian@lbl.gov