Boise State University
ScholarWorks

Electrical and Computer Engineering Faculty Publications and Presentations

Department of Electrical and Computer Engineering

1-1-2004

Protein Family Classification Using Structural and Sequence Information

Jennifer A. Smith
Boise State University

This document was originally published by IEEE in *Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*. Copyright restrictions may apply. DOI: 10.1109/CIBCB.2004.1393950

Protein Family Classification Using Structural and Sequence Information

Scott F. Smith, *Member, IEEE*

***Abstract*—Protein family classification usually relies on sequence information (as in the case of hidden Markov models and position-specific scoring matrices) or on structural information where some sort of average positional error between the atomic locations is used. The positional error method requires that the structure of all the proteins to be classified is known. Sequence methods have the advantage that a much larger number of proteins can be classified (since far more sequences are know than structures). However, sequence methods discard a large amount of useful information contained in the structures of the subset of proteins in the family for which structures are known. A protein family classification system is presented which uses both structural and sequence information and combines this information in a way consistent with fuzzy systems theory. The non-linear fuzzy-theory-based method is found to perform better than either an equally-weighted linear combination of the sequence and structural information or the sequence information alone.**

***Index Terms*—Biological sequence analysis, computational molecular biology, fuzzy systems, proteins.**

I. INTRODUCTION

The classification of proteins into families is useful because it can suggest possible functions and structure for proteins where these are unknown. A number of protein classification databases exist, including the *Structural Classification of Proteins* (SCOP) [1], *Class, Architecture, Topology, and Homologous Superfamily* (CATH) [2], and *Protein Family* (Pfam) [3] databases. Methods used to generate these classifications include sequence-only automated methods such as profile hidden Markov models (profile HMM) [4] and position-specific scoring matrices (PSSM) [5] as well automated structural alignment and hand curation.

When the sequence of a new protein to be classified is known, but the three-dimensional structure is not, then comparison of the new protein with profile HMM or PSSM of existing protein families is the normal course of action. These methods often work since most proteins contain conserved regions that are similar to other proteins [6]. These conserved regions are normally in the hydrophobic core of the protein and are mostly composed of alpha helices and beta sheets. Amino acids on the surface of proteins are much more variable and form loops that connect the alpha helices and beta sheets. Sequence-only methods like the HMM and PSSM find conserved and non-conserved regions directly from the sequence data. The probabilistic model discovers the degree of conservation from the observed sequence data, but does not use any structural information to determine locations where conservation should be more likely due to being in the protein core. If the number of known proteins in the family is very large, then the observed sequences are a good measure of whether a given position is conserved or not. If very few sequences in a family are observed, then the degree of conservation at a location within the alignment is harder to estimate.

In this paper, structural data from the subset of protein family members with known structure is used to estimate the degree of membership of a protein sequence location in the set of conserved sequence locations using the structural data. All family members are used to estimate the degree of membership of sequence locations in the set of conserved sequence locations using sequence data. A combined degree of membership estimate at each sequence location is found using standard fuzzy logic operations. This combined degree of membership estimate is then used to weight the scores of a PSSM.

Section II details how the fuzzy-theory-based estimate of the degree of conservation at each residue position is obtained and also describes two alternative estimates of degree of conservation to be tested. The method used to score a new protein sequence for which structural information is not known is discussed in Section III. The Monte Carlo simulation method to determine the level of significance of the score is presented in Section IV. Section V gives a specific protein domain family example, the TPR domain from the Pfam database and compares the sensitivity of the three conservation estimates. The specificity of the three estimates is examined in Section VI. The specificity is examined in terms of proteins from nineteen other families not matching the three models determined for the TPR family. Sensitivity for three other protein families is examined in Section VII. The computational complexity of the algorithm is addressed in Section VIII. Conclusions are drawn in Section IX.

Manuscript received May 18, 2004. This work was supported in part by the U.S. National Institutes of Health under Grant P20RR16454.

The author is with the Department of Electrical and Computer Engineering at Boise State University, Boise, ID 83725 USA (phone: 208-426-5743; fax: 208-426-2470; e-mail: sfsmith@boisestate.edu).

0-7803-8728-7/04/$20.00 ©2004 IEEE

Authorized licensed use limited to: Boise State University. Downloaded on April 28, 2009 at 11:53 from IEEE Xplore. Restrictions apply.

II. DEGREE OF CONSERVATION ESTIMATES

A multiple alignment is first performed on all of the sequences known to be in the family and the sequence that is to be tested for family membership. The sequence to be tested is then removed from the multiple alignment for the purposes of estimating the degree of conservation and the position-specific scoring matrix.

A. Conservation Estimate from Sequence Information

The degree of conservation from sequence data is estimated by placing the residues at a given multiple. The degree to alignment sequence position into one of four groups. These groups are hydrophobic (A, V, L, I, M, F, and P), charged (D, E, R, and K), polar (S, T, Y, H, C, N, Q, and W), and glutamine (G). The largest of the four calculated fractions becomes the estimate of the degree of conservation from sequence. The fractions will not necessarily add up to one, since gaps in the multiple alignment do not appear in any group.

B. Conservation Estimate from Structural Information

The spatial locations of the alpha carbon atoms in the protein backbone are obtained from *Protein Data Bank* (PDB) [7] files for the subset of proteins where structure is known. Those amino acid positions that are near the surface are then estimated using a simple algorithm that determines the furthest extent of the protein in twenty-six different directions. These directions are the major axes (six directions), all pairs of major axes (twelve directions), and all triplets of major axes (eight directions). These positions are then extended by two amino acid positions in either direction along the amino acid chain. The fraction of sequences that are non-surface at a given position becomes the conservation estimate from structural information. Gaps in the multiple alignment are treated as if they were surface positions.

C. Combined Degree of Conservation Estimate

We would like to combine the estimated level of conservation from the sequence information with that obtained from the structural information. One way to do this would be to take a weighted average of the two estimates. However, there is no clear way to choose the weights to be placed on the two initial estimates. It could be argued that the weights should be estimated from data, where conservation estimates from many protein families are made and the reliability of the two estimates assessed in each case. The weights are then chosen as inverse to the measured reliability. This implicitly assumes that reliability of the estimates are similar between families. However, large variations in the fraction of proteins in a family with known structure and the alpha helix versus beta sheet structures of the hydrophobic core may mean that the reliability of the two estimates is much different for different families.

Instead, a method consistent with fuzzy logic theory [8] is used. We wish to define an amino acid location as conserved if either the structural data indicates the location is very conserved, or the sequence data indicate the location is very conserved, or both sets of data indicate that the location is at least moderately conserved.

Formally, we define Q as the set of sequence locations that are conserved based on the sequence data. The set T is the set of sequence locations that are conserved based on the structural data. The set C will be the set of sequence locations that are conserved based on the combined data which a location x is a member of Q, T, or C respectively is given by $mQ(x)$, $mT(x)$, and $mC(x)$. The sets of locations that are very conserved based on sequence or structural data are VQ and VT respectively. Using the concentration operation the membership function of the sets VQ and VT can be obtained as

$$mVQ(x) = [mQ(x)]^2 \text{ and } mVT(x) = [mT(x)]^2. \qquad (1)$$

The statement that a location should be either very conserved based on sequence data, or very conserved based on structural data, or at least somewhat conserved in both can be written as

$$C = VQ \cup VT \cup [Q \cap T]. \qquad (2)$$

The associated membership function for C can be calculated from

$$mC(x) = \max\{mVQ(x), mVT(x), \min[mQ(x), mT(x)]\}. \qquad (3)$$

In order to avoid having either the sequence or the structure dominate the combined membership function, mVQ, mVT, mQ, and mT are all normalized to have a mean of 1.0 before doing the above calculation. The mC result is also normalized to have a mean of 1.0 after the calculation so that the use of mC as a weight can be compared to using the equally weighted average $(mQ + mT)/2$. In what follows, the membership function values for each sequence position will be used as weights for the relative importance of matching a position in the position-specific scoring matrix (PSSM).

III. SCORING A NEW PROTEIN SEQUENCE

To test if a new sequence is similar to the existing family, the new sequence is scored by comparing the residue observed at each position in the aligned new sequence to a weighted measure of the frequency of occurrence of that residue in that position among the known family members.

First, the number of non-gaps in the family members at each position is counted. For each of the twenty possible amino acids, the number of observations of that amino acid at the position is then counted. A pseudo-count of one is

Authorized licensed use limited to: Boise State University. Downloaded on April 28, 2009 at 11:53 from IEEE Xplore. Restrictions apply.

added to each of the counts such that a score of zero at any position will not dominate the estimate. The unweighted PSSM is the number of observed amino acids divided by the number of non-gaps.

The weighted PSSM is formed by multiplying each location in the unweighted PSSM by the combined degree of conservation from part II above. The log of each element of this PSSM is then taken to solve computing precision problems. The score of a new sequence is simply the sum of the individual PSSM values at each location corresponding to the observed residue in the new sequence. Three such PSSM are generated, one each using the weights mC, $(mQ + mT) / 2$, and a constant weight of 1. The first and second will allow comparison of using the non-linear fuzzy-based estimate of conservation to a simple linear estimate where both use structural and sequence information. The third PSSM uses only sequence information and can be used detect whether inclusion of structural information has any benefit at all.

IV. SIGNIFICANCE OF THE SCORE

In order to determine the significance of the scores resulting from the three PSSM when compared to the null hypothesis that the score was generated by chance, Monte Carlo simulations are run with 1000 reshuffled versions of the sequence under test. The mean score generated from these reshuffled versions is taken as the null hypothesis score. Since the scores are generated in terms of logs, the difference between the unshuffled test sequence score and the null hypothesis score is a measure of the significance level of the score. A base-two log is taken when generating the PSSM, so the units of significance are bits. A significance measure of 10.0 will therefore imply that the test sequence match is 1024 times (2^{10}) more likely than pure chance.

It should be noted that by using a reshuffled version of the test sequence, we have eliminated the possibility of matching based on the test sequence and the family having the same order-independent ratios of amino acid occurrences. However, this information itself may have some (but probably not much) explanatory power as to whether the test sequence is a member of the family. As such, this method generates a conservative (high) estimate of the null hypothesis score.

V. TPR DOMAIN EXAMPLE

To test the performance of the classification scheme developed in this paper, the significance levels of the score using fuzzy weighting is compared to the significance levels using the equally-weighted average and to no weighting.

As an example domain, the *tetratrico peptide repeat* (TPR) domain from the Pfam database was used. Reasons for choosing this family include the fact that it is in the "top-twenty" list on the Pfam site of most-frequently occurring protein families and that it has an adequate number of sequences with know structure (twelve). There are a total of 575 sequences in the "seed" family of hand-curated sequences, including sequences of both known and unknown structure. There is an average of 18% sequence identity among the sequences which places this domain in the "twilight zone" of remote homologs [9] that are difficult to classify. The multiple alignment for this domain is 34 residues long.

Table I shows the twelve sequences out of the total 575 sequences for which the three-dimensional structure is known. The first column shows the Swiss-Prot name that is used for the sequence within the Pfam multiple alignment file (the family as a whole has Pfam identifier PF00515). Several of the domain sequences actually come from the same polypeptide molecule, where this domain appears more than once along the molecule length. As a result, there are in fact only six different proteins (ncf2_human, fkb5_human, ppid_bovin, ppp5_human, pex5_human, and iefs_human). The second column of Table I shows the Protein Data Bank (PDB) name for the same protein. There are actually seven PDB proteins listed (1hh8, 1kt0, 1ihg, 1a17, 1fch, 1elw, and 1elr) since iefs_human corresponds to 1elw in one case and 1elr in two other cases. The two PDB entries are just slightly different versions of the same protein. The third column of Table I shows the residue positions of the domain within the total protein. These residue positions are the same for the Swiss-Prot and PDB proteins used here, but in general this might not be the case.

A. Degree of Conservation Estimate

A test sequence is selected at random and removed from the set of sequences used to estimate the model of the protein domain family. The test sequence is always chosen from the 563 sequences that do not correspond to known three dimensional structures since there are so few known structures.

The coordinates files for the seven PDB proteins are obtained from the PDB database and the relative three dimensional locations of the alpha carbon atoms of each residue are extracted. The surface residue sequence letters

TABLE I
TPR DOMAINS WITH KNOWN STRUCTURE

Swiss-Prot Name	PDB Name	Residues
ncf2_human	1hh8	71-104
fkb5_human	1kt0	317-350
fkb5_human	1kt0	351-384
ppid_bovin	1ihg	273-306
ppid_bovin	1ihg	307-340
ppp5_human	1a17	28-61
ppp5_human	1a17	96-129
pex5_human	1fch	451-484
pex5_human	1fch	485-518
iefs_human	1elw	4-37
iefs_human	1elr	225-258
iefs_human	1elr	300-333

Authorized licensed use limited to: Boise State University. Downloaded on April 28, 2009 at 11:53 from IEEE Xplore. Restrictions apply.

are then converted to upper case and the interior residue sequence letters to lower case using the Matlab code available at [10]. The twelve sequences corresponding to known structures in the Pfam multiple alignment file are then moved to the top of the file and are converted to upper or lower case based on the results of the previous step. A second Matlab program (also available at [10]) is used to score the test sequence and calculate level of significance.

Figures 1 and 2 show the estimated conservation weights using *mC* (fuzzy), and the linear combination of *mQ* and *mT* respectively. The fuzzy-based conservation estimator places significantly more weight on four of the amino acid positions (20, 27, 28, and 30) which have consensus residues of A, A, L, and L respectively.

B. Protein Sequence Scores

The number of occurrences of each of the twenty possible amino acids is counted at each of the 34 multiple alignment positions. A pseudocount of one is added to each count to avoid taking a log of zero and having the non-observance of a particular residue at a particular position absolutely rule out accepting a sequence with that residue at that position. The resulting counts are divided by the total number of sequences (574) and a base-two log taken. Since there may be gaps in the multiple alignment, the sum of the amino acid counts might not equal the total number of sequences. The result is the unweighted PSSM.

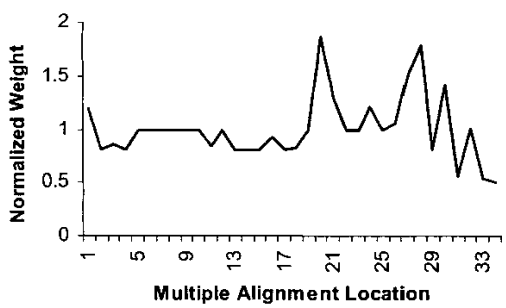

Fig. 1. Conservation weights using fuzzy estimator

Fig. 2. Conservation weights using linear estimator

The unweighted PSSM in now multiplied by the weights for each of the 34 positions determined from the conservation estimates above. The test sequence is then scored against each of the three PSSM. This is done by adding PSSM values for each position corresponding to the residue at that position in the test sequence.

C. Significance of the Scores

The process in parts A and B above is repeated 20 times with a new test sequence randomly selected from the 563 sequences of unknown structure each time. The resulting three PSSM are slightly different each time due to the removal of a new test sequence and the reinsertion of the old test sequence, but with 575 sequences total the difference is very small. Since there is a new test sequence, the estimate of the score under the null hypothesis is different each time.

For each of the 20 runs, the significance level (in bits) generated using the fuzzy and linear estimates of conservation and without weighs is calculated. The significance levels from each of the 20 runs is then used to calculate the sample mean and variance. The sample means are reported in Table II. The non-linear fuzzy combination appears to outperform the other methods.

To determine if the significance level results could reasonably be expected from pure chance a t-test of the mean difference between significance levels is undertaken [11]. The mean difference between fuzzy and linear and the mean difference between fuzzy and unweighted is examined. In each case the null hypothesis is that the mean difference is zero and the alternative hypothesis is that the mean difference is positive. The null hypothesis can be rejected at the 99% confidence level if the t-statistic is greater than 2.54, where the number of degrees of freedom is 19. Table II shows that there is at least a 99% chance that the fuzzy estimator returns a score significance higher than the other estimators in the population.

The data used to generate the results in Table II are shown in Table III. The first column shows the protein name of the randomly selected sequence and the residue location within the protein. The rightmost three columns show the significance levels of the scores using the fuzzy, linear, and unweighted (flat) PSSM respectively.

TABLE II
PERFORMANCE ON TPR DOMAIN

Parameter	Value
Domain Structures	12
Domain Sequences	575
Test Sequences	20
Mean Significance Difference:	(units of bits)
Fuzzy-Linear	3.94
Fuzzy-Unweighted	8.84
t-Statistics:	
Fuzzy-Linear	7.37
Fuzzy-Unweighted	10.52
99% Confidence	2.54

Authorized licensed use limited to: Boise State University. Downloaded on April 28, 2009 at 11:53 from IEEE Xplore. Restrictions apply.

TABLE III
INDIVIDUAL TPR SIGNIFICANCE LEVELS (BITS)

Sequence/Residues	Fuzzy	Linear	Flat
cya3_rhime/455-488	93.32	89.97	82.72
q43468/415-448	91.77	89.55	92.10
ppp5_rat/96-129	90.19	89.43	84.32
o51228/809-842	88.24	82.90	81.18
pex5_pican/450-483	100.63	94.82	94.32
o82039/334-369	111.60	103.66	99.11
solr_cloab/133-166	88.43	85.49	78.97
nuc2_schpo/499-532	94.81	93.16	90.90
p90647/316-349	100.06	94.29	85.83
c27_yeast/540-573	101.43	101.97	99.08
p74123/124-157	98.46	96.60	90.75
p74321/193-226	93.04	89.68	81.13
yct3_marpo/72-105	101.28	99.40	91.39
rapc_bacsu/223-256	94.07	90.82	84.20
o26176/118-151	106.53	99.42	95.96
ogt1_rat/215-248	90.71	86.06	80.06
ppp5_mouse/28-61	102.26	93.45	88.85
ttc1_human/189-222	83.74	79.35	75.45
klc1_human/377-410	86.40	82.68	74.34
ctr9_yeast/218-251	98.79	94.24	88.26

TABLE V
INDIVIDUAL NON-TPR SIGNIFICANCE LEVELS

Sequence (Size)	Family	Fuzzy	Lin.	Flat
env_hv1b1(856)	gp120	14.13	16.40	14.81
tsh_drome(993)	zf-c2h2	9.37	8.91	9.31
q9h069(225)	lrr	14.40	16.71	17.13
ym40_marpo(502)	rvt	11.23	11.44	12.47
pol_omvvs(1086)	rvp	15.40	16.59	14.90
cyb_ascsu(365)	cytochrom_b_n	16.04	18.40	17.99
q9zem4(1049)	wd40	13.64	18.65	17.71
q01484(3924)	ank	9.60	11.27	10.11
cox1_hanwi(535)	cox1	12.68	14.36	13.38
nu2m_apili(333)	oxidored_q1	10.94	17.01	17.32
petd_chleu(160)	cytochrom_b_c	17.05	16.90	14.85
nike_ecoli(268)	abc_tran	16.81	16.97	15.35
mk04_hum(557)	pkinase	19.72	19.28	18.34
rbl_antsp(488)	rubisco_large	13.38	14.03	14.61
rbl_cyapa(475)	rubisco_large_n	13.59	14.95	12.97
q31377(246)	ig	10.67	14.15	17.00
o80524(705)	ppr	20.06	23.96	22.91
o57059(986)	rvt_thumb	11.52	11.27	10.30
polg_hcvbk(3010)	hcv_ns1	10.57	18.67	17.54

VI. TPR MODEL APPLIED TO NON-TPR PROTEINS

A random selection of one protein sequence from each of the nineteen other "top twenty" families in the Pfam database is selected and scored against the three TPR models of the previous section. Every possible subsequence of 34 characters from the entire protein sequence (not just the subsequence that forms the domain of the family) is scored against the TPR PSSM models. The highest significance score from any 34-character subsequence is retained.

Since the highest score significance is retained for each protein, the scores are expected to follow an extreme value distribution [9] where the expected score significance increases with the log of the number of residues in the protein. The maximum score significances for each protein are divided by the log of the sequence size to normalize this size-dependent effect.

The difference between the fuzzy-based significance and the linear-combination-based significance is taken as well as the difference between the fuzzy-based significance and the unweighted significance. The mean significance differences are reported in Table IV. It is desirable that these score significances be low since they represent rejection of non-family proteins. In both cases, the mean significance of the fuzzy-based score is lower. The t-statistics show that the mean difference between the fuzzy-based score significance and the non-fuzzy-based score significance is negative at the 97.5% confidence level. The fuzzy-based conservation weights therefore result in lower rates of false-positives as well as lower rates of false-negatives (Section V).

TABLE IV
PERFORMANCE ON NON-TPR DOMAIN

Parameter	Value
Test Sequences	19
Mean Significance Difference:	(units of bits)
Fuzzy-Linear	-2.06
Fuzzy-Unweighted	-1.48
t-Statistics:	
Fuzzy-Linear	-3.80
Fuzzy-Unweighted	-2.33
97.5% Confidence	2.10

The 19 randomly selected sequences are shown in Table V along with their size (total number of residues) and the Pfam family to which they have been assigned. The rightmost three columns of the table show the size-normalized significance scores obtained from each of the three PSSM models. The family with the worst rejection performance (PPR) is the most closely related of the 19 families tested to the TPR family. The significance scores are positive in all cases. This is to be expected since existing biological proteins are more closely related than randomly generated amino acid sequences. Randomly generated amino acid sequences should have a mean significance score of zero by definition. However, randomly generated proteins on average are not biologically stable and many proteins from different families may be very remotely evolutionarily related.

VII. OTHER DOMAIN FAMILY TARGETS

The characteristics of domains in the Pfam top twenty are listed in Table VI. The number of sequences and the number of sequences with known structure within the hand-curated "seed" family are shown in columns two and three. The four families with ten or more known structures are marked with an asterisk (LRR, WD40, Ank, and TPR). There is potential difficulty in classifying new sequences into these four

Authorized licensed use limited to: Boise State University. Downloaded on April 28, 2009 at 11:53 from IEEE Xplore. Restrictions apply.

TABLE VI
PFAM TOP-20 DOMAIN CHARACTERIZATION

Family Name	Number Sequence	Number Struct.	Percent Identity	Average Length
gp120	24	0	56	154
zf-c2h2	197	8	36	23
lrr	2651	15*	26	24
rvt	177	0	70	161
rvp	53	2	86	94
cytochrom_b_n	8	0	69	152
wd40	1923	23*	20	39
ank	1181	50*	27	30
cox1	24	1	47	227
oxidored_q1	33	0	29	221
cytochrom_b_c	9	0	74	89
abc_tran	63	1	26	184
pkinase	67	6	23	219
rubisco_large	17	2	79	282
rubisco_large_n	17	2	83	117
tpr	575	12*	18	34
ig	91	6	21	64
ppr	560	0	20	33
rvt_thumb	42	0	88	50
hcv_ns1	10	0	45	74

families since they all have low sequence identity (26, 20, 27, and 18 percent) and are rather short (24, 39, 30, and 34 amino acids) as can be seen from the last two columns of the table.

The analysis from Section V applied to the TPR domain is also applied to the LLR, WD40, and Ank domains, with the results shown in Table VII. With the exception of the WD40 fuzzy versus unweighted case, the fuzzy gives a higher mean significance difference at the 99% confidence level. The WD40 fuzzy versus unwieghted case is statistically significant at the 95% confidence level.

VIII. PERFORMANCE

The algorithms were run on an 800 MHz Pentium III using Matlab version 6 release 13. There are two tasks for which performance measures are of interest, the CPU time required to obtain the score significance level given that the PSSM for a family has already been calculated and the CPU time required to obtain the PSSM for the family in the first place. The performance of these algorithms probably could be increased significantly since no attempt has been made to optimize the code and recoding in a language such as C is likely to speed up the calculations.

For a test sequence with 60 residues and using 1000 reshuffles of the test sequence to obtain a score significance level 4.05 seconds of CPU time was required. More than 99% of this time is spent calculating the scores of the 1000 reshuffled versions of the test sequence. If performance of the algorithm is an issue, analysis of the minimum number of reshuffles needed to get an acceptable result should be undertaken. The required CPU time is very close to linear in both number of reshuffles and test sequence length. There are 7459 families in Pfam as of June 2004. Estimating that about one fourth of the Pfam entries have enough known-structure members for the method of this paper and that the average family has a PSSM with about 200-300 residues means that about eight hours of CPU time is needed to search all possible families. Using a faster computer, some reduction in number of reshuffles, and recoding in a more efficient language should allow this time to be reduced to well under an hour.

Calculating the PSSM for a protein family with 1180 sequences of aligned length 60 and with 36 of the structures known requires 1.6 seconds of CPU time. For a known structure of length 508 residues, it takes 0.31 seconds of CPU time to find the surface residues and the required time is nearly linear in number of residues. Finding the surface residues for the example above would therefore take about 1.3 seconds of CPU time for a total of 2.9 seconds to find the PSSM. Recalculating the PSSM for all the possible Pfam entries would therefore take a fraction of a day.

IX. CONCLUSIONS

The introduction of non-linearity into the estimation of conservation weights for PSSM scoring seems to improve performance versus linear estimation or non-weighted scoring. The fuzzy-based conservation estimator generates a PSSM score with greater sensitivity (lower false negative) and greater specificity (lower false positive) than the equally-weighted linear conservation estimator or unweighted scoring when applied to the TPR, LLR, WD40, and Ank protein families.

While the results of adding structural information to the sequence information in this manner are promising, it is only applicable to those protein families that have a significant number of members with structures that have been determined. As the number of structures in the Protein Data Bank increases the desirability of including this structural information in models for classifying new protein sequences into protein families will increase.

TABLE VII
PERFORMANCE ON LLR, WD40, AND ANK

Parameter	Value
Mean Significance Difference:	(units of bits)
LLR Fuzzy-Linear	29.16
LLR Fuzzy-Unweighted	28.60
WD40 Fuzzy-Linear	25.44
WD40 Fuzzy-Unweighted	3.96
Ank Fuzzy-Linear	20.46
Ank Fuzzy-Unweighted	3.89
t-Statistics:	
LLR Fuzzy-Linear	21.36
LLR Fuzzy-Unweighted	11.43
WD40 Fuzzy-Linear	19.21
WD40 Fuzzy-Unweighted	2.08
Ank Fuzzy-Linear	17.19
Ank Fuzzy-Unweighted	3.04
99% Confidence	2.54

Authorized licensed use limited to: Boise State University. Downloaded on April 28, 2009 at 11:53 from IEEE Xplore. Restrictions apply.

REFERENCES

[1] A. Murzin, S. Brenner, T. Hubbard, C. Chothia, "SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures," *J. Mol. Biol.*, vol. 247, pp. 536-540, 1995.

[2] C. Orengo, A. Michie, S. Jones, D. Jones, M. Swindells, and J. Thornton, "CATH- A Hierarchic Classification of Protein Domain Structures," *Structure*, vol. 5, pp. 1093-1108, 1997.

[3] A. Bateman, L. Coin, R. Durbin, R. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. Sonnhammer, D. Holme, C. Yeats and S. Eddy, "The Pfam Protein Families Database," *Nucleic Acids Res.*, vol. 32, pp. D138-D141, 2004.

[4] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis*, Cambridge University Press, 1998.

[5] D. Mount, *Bioinformatics*, Cold Spring Harbor: CSHL Press, 2001.

[6] C. Branden and J. Tooze, *Protein Structure*, 2nd ed., London: Garland, 1999.

[7] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, P. Bourne, "The Protein Data Bank," *Nucleic Acids Res.*, vol. 28, pp. 235-242, 2000.

[8] N. Kasabov, *Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering*, Cambridge: MIT Press, 1996.

[9] C. Orengo, "Sequence Comparison Methods," in *Bioinformatics: Genes, Porteins, and Computers*, C. Orengo, D. Jones, and J. Thornton, Eds. Oxford: BIOS Scientific Publishers, 2003.

[10] S. Smith, *Matlab Code for Protein Family Classification Using Structural and Sequence Information*, http://coen.boisestate.edu/ssmith/biohw/compcode.

[11] G. Keller and B. Warrack, *Statistics for Management and Economics*, 4th Ed., p. 376, Pacific Grove: Duxbury Press, 1997.

Authorized licensed use limited to: Boise State University. Downloaded on April 28, 2009 at 11:53 from IEEE Xplore. Restrictions apply.