Identification of Sequence Patterns with Profile Analysis

Short Title: Profile Analysis

Michael Gribskov
gribskov@sdsc.edu
(619) 534-8312
(619) 534-5117 FAX

Stella Veretnik
veretnik@sdsc.edu
(619) 534-8317

San Diego Supercomputer Center
P.O. Box 85608
San Diego CA 92186-9784
(619) 534-5152 FAX

Introduction

Ten years ago, it was rare for more than one or two sequences belonging to a homologous family to be known. Today, the situation is dramatically different and we are quickly approaching the time when it will be rare that a newly determined sequence does not fall into a known family. This amazing growth of molecular sequence data has brought us to the paradoxical point where a database search with a new sequence now may reveal so many significantly related sequences that it becomes difficult to decipher what they have in common. The increasingly overwhelming nature of sequence data has led to a number of efforts to organize information into libraries of motifs (e.g., PROSITE [1], BLOCKS [2],and ProDom [3]), and more recently into more sophisticated mathematical models of protein families based on hidden Markov models [4,5]. At the same time, improvements in our ability to determine realistic three-dimensional models of proteins based on the structures of homologous proteins has lead to an increased interest in the conserved regions of families of sequences because the highly conserved regions correspond to structurally conserved regions. This link between conserved regions in sequence families and the core of three-dimensional structures makes methods, such as profile analysis, that allow the information present in sequence alignments of protein families to be used in homology modeling increasingly important.

The idea of a profile is straightforward and easy to understand. The profile is a weight matrix that, for each position in a group of aligned sequences, assigns a score for each of the twenty possible amino acid residues. At its simplest, the profile can be thought of as merely a convenient data structure capable of encoding the character of conserved residues seen in a group of related sequences. At a more sophisticated level, however, each profile can be seen as mathematical model for a group of protein sequences. This model is more complex than a simple weight-matrix model in that it contains position specific information on insertions and deletions in the sequence family, and is quite closely related to the one employed in hidden Markov models for sequences. The profile approach has turned out to be quite flexible with applications ranging from

describing DNA sequence motifs, to the characterization of protein families, and to the mapping of sequences onto three dimensional structures [6,7,8].

Methods

Profile Analysis

The profile, figure 1, is a two dimensional weight matrix in which the rows correspond to aligned positions in a group of sequences, and the columns correspond to each of the twenty possible amino acid residues (or four DNA bases). The profile uses a similarity based scoring system where positive values indicate that the residue represented by the column in which the value occurs is similar to the corresponding residues in the aligned sequences, and negative values indicate dissimilarity. Profiles differ from generic weight matrices in having two additional columns that specify position specific weights for gap penalties. The two additional columns represent weights on the gap opening penalty and the gap extension penalty.

<Place Figure 1 here>

Profiles can be matched with sequences using an extension of standard dynamic programming sequence alignment techniques, usually the algorithm of Smith and Waterman [9]. A linear gap penalty (often called an affine gap penalty) consisting of a length independent (gap opening) and length dependent (gap extension) term is widely considered to be the most useful for sequence alignments and the two gap weights included in the profile allow these terms to be separately modified. A formal description of the alignment of profiles and sequences has been presented [10] and will not be repeated here.

The Profile Analysis package of programs [11] provides a suite of tools for creating profiles and matching them with sequences. These programs have been described in detail in earlier papers and we will describe them only briefly here:

- PROFILEMAKE [10,12] is used to create profiles from groups of aligned sequences.

- PROFILEGAP [10,13] is used to align a profile and one or more sequences. A useful newly added function is the ability to produce a multiple alignment of a group of sequences to a profile, in addition to the pairwise alignments available previously.
- PROFILESEARCH [10,14] uses a profile as a query in a database search, normalizes the results for systematic dependence on length, and converts the scores to Z scores (standardized scores). The previously independent PROFILENORMAL program has been incorporated into PROFILESEARCH. A supercomputer implementation of PROFILESEARCH called PROFILE-SS is available [15] for the CRAY C90.
- PROFILESCAN [16] compares a single sequence to a library of statistically characterized profiles. The current library consists of over six hundred protein sequence motifs.

Average Profiles

We refer to the original method for calculating profiles as the average method. Briefly, a single fixed scoring table, e.g., the PAM 250 table [17] , is used as the basis of the profile. This table can be thought of as specifying 20 model residue frequency distributions, one for each possible ancestral residue. The model distributions are combined into a mixture distribution with the components weighted by the relative frequencies in the observed distribution:

$$\text{Profile}_{ij} = \sum_{k=1}^{20} f_{ik} M_{jk} \qquad (1)$$

where f_{ik} is the relative frequency of residue k at position i in the aligned sequences, and M_{jk} is the comparison score for residues j and k in the basis scoring table. Note that the relative frequencies f_{ik} may be adjusted in various ways to account for sampling error or bias (see Gribskov et al.[10], and below).

Evolutionary Profiles

We have recently developed a finite mixture method based on the Dayhoff model of protein evolution [17] for the calculation of profiles. We refer to this method as the evolutionary profile method to distinguish it from the earlier average profile method. The evolutionary profile method models each position in the observed group of sequences as arising from one or more ancestral residues each possibly at a different evolutionary distance. In adopting this approach we are attempting to explicitly include biologically relevant prior information about the rate and kind of change occurring at each position in a homologous family.

The evolutionary profile method requires two steps at each position in the aligned sequences: first, the Dayhoff evolutionary model is used to generate a series of model distributions for each of the twenty possible ancestral residues at various evolutionary distances (typically 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, and 2048 PAM distances). The evolutionary distance that minimizes the cross entropy, H, of the model and observed distributions is chosen for each possible ancestral residue. A different evolutionary distance is fit for each possible ancestral residue giving twenty model distributions in all.

$$H = -\sum_{a=1}^{20} f_a \ln p_a \qquad (2)$$

Where f_a are the observed residue frequencies and p_a are the predicted frequencies in the model distribution at a specific evolutionary distance.

In the second step, the twenty model distributions determined above are mixed based on the probability that each one could give rise to the observed frequency distribution. This is done by determining a weight (mixture coefficient) for each of the twenty model distributions such that the weight corresponds to the extent to which the model predicts the observed distribution.

The probability of the model distribution, M_a, for a given ancestral amino acid residue, a, giving rise to the observed residue frequency distribution F is given by

$$P(M_a|F) = \frac{P(M_a) \times P(F|M_a)}{\sum_{a=1}^{20} P(M_a) \times P(F|M_a)} \quad (3)$$

where $P(M_a)$ is the prior distribution for the amino acid residues, normally the amino acid residue frequencies in the database, and

$$P(F|M_a) = \prod_{a=1}^{20} p_a^{f_a} \quad (4)$$

Mixture coefficients, W, are given by $W_a = P(M_a|F) - P(M_{random}|F)$ where M_{random} is the residue frequency distribution of random sequences, i.e., the amino acid residue frequencies in the database. Typically only a few residues will have weights appreciably greater than zero (see figure 2). Because M_a at very long PAM distances is identical to M_{random}, the weights are guaranteed to be positive. Finally the profile is calculated as the log-odds ratio for the weighted sum of the mixture components:

$$\text{Profile}_{ij} = \ln\left(\sum_{a=0}^{20} \left(W_{ai} p_{aij} \right) \Big/ p_{\text{random } j} \right) \quad (5)$$

W_{ai} is the weight on the ancestral residue, a, at position i, and p_{aij} is the frequency of residue j in the ancestral residue frequency distribution a at position i.

<Place Figure 2 here>

Sequence Weighting

Families of sequences are nearly always highly biased. This is primarily due to bias in our selection of organisms to sequence, typically focusing on mammals, yeast, E. coli, and drosophila. In any case, it is common to find in a group of sequences that several are nearly identical, and several others are as little as 20% identical when aligned. It is clear in this case that each of the nearly identical sequences contributes much less information than each of the 20% identical sequences. Weighting procedures seek to correct for this sampling bias, hopefully correcting the

observed residue counts so that they correspond to a random sample. In this work, sequences have been weighted using the approach of Felsenstein [18]. However, rather than basing the weights on a completely resolved phylogenetic tree, the weights are based on an approximate tree where the distances are the percentage of differing residues in pairwise alignments. Briefly, a single-linkage multifurcating tree is constructed in which branches are joined at nodes representing discrete distance thresholds. For instance, all sequences with less than 10% differing residues are joined at a single node. This node is joined with all sequences with less than 20% differing residues at the next node, and so on. Because these trees are approximate, they are robust to small errors in alignments and insensitive to the fine points of tree topology.

Evaluating Matching with ROC Analysis

The receiver operating characteristic (ROC) is a widely used technique for evaluating the performance of clinical tests and treatments (for a review see Zweig and Campbell [19]). ROC analysis has several advantages over other techniques. The ROC is a function of both the sensitivity of an assay (what fraction of the true positives are detected) and the specificity of the assay (how well are the true positives separated from the true negatives). One of its major advantages is threshold independence; the entire distribution of scores is examined rather than just scores over an arbitrary significance threshold.

<Place Figure 3 here>

ROC analysis involves the construction of an ROC plot (fig 3). The plot is constructed by examining each observation, in this case, each sequence in the results of a database search, and plotting the fraction of true positives (homologous family members) and true negatives (unrelated sequences) with equal or higher scores on the ordinate and abscissa respectively. The area under the curve of the ROC plot measures the probability of correct classification and is a simple statistic that can be used to compare searches using different query sequences or conditions (higher values indicate better performance in detecting the homologous family). We have recently introduced the

ROC_{50} for the evaluation of sequence database searches [20,21]. The ROC_{50} is the area under an ROC curve where the list of results is truncated after observing 50 negative sequences, i.e., the number of true negatives is exactly 50.

Results

4Fe-4S Ferredoxins

The 4Fe-4S ferredoxins are small proteins involved in electron transfer (for a recent review see Beinert [22]). Ferredoxin like molecules also function in photosynthesis, and are found in a variety of enzymes involved in oxidation-reduction reactions, e.g., succinate, fumarate, and glycerol-3-phosphate dehydrogenases, dimethyl sulfoxide reductase, formate hydrogenlyase, and sulfite reductases. These ferredoxins bind a 4Fe-4S cluster at a highly conserved 12 residue sequence. The core of the conserved region has the consensus pattern C-X-X-C-X-X-C-X-X-X-C-[PEG] and insertions or deletions are not usually required to align this region. In Swiss-Prot release 31.0 there are 134 4Fe-4S ferredoxins (including the "false negative" members of the family not detected by the PROSITE signature). Most of the members of this family have two copies of the characteristic 12 residue repeat and bear two 4Fe-4S centers. For the work described here we generated profiles for the 19 residues beginning two residues before the first conserved cysteine and ending six residues after the last conserved cysteine.

<Place Table I here>

Table I shows a comparison of the efficacy of profiles made by the average and evolutionary methods in detecting the 4Fe-4S family. Note that because the sequences have been limited to the most highly conserved region, we are not taking full advantage of the ability of the profile to distinguish conserved and unconserved regions. Similarly, because matching to this region does not require gaps, we obtain no advantage from the position specific gap penalties that can be encoded in a profile, an important feature when matching to distantly related sequence families (see

Gribskov et al.[10]). The differences between the average and evolutionary profile methods therefore correspond only to their respective abilities to generalize the sequence pattern typical of the family from the observed set of sequences. The evolutionary profile consistently performs better than the average profile. When all sequences are included, the evolutionary profile has only about one third the average classification error (1-ROC_{50}) as the average profile, a small but important difference.

<Place Figure 4 here, or facing following section>

ATP Dependent RNA Helicases

The members of this family are involved in ATP dependent unwinding of nucleic acids. This family is also known as the "DEAD helicase" family due to the presence of a highly conserved sequence [LIVM]-X-X-D-E-A-D-[RKEN] at what is thought to be part of the ATP binding site. The proteins comprise a number of conserved blocks distributed across the length of the sequences. As an example, we have focused on the conserved block containing the DEAD signature (fig 4). As can be seen, this conserved block requires that some insertion/deletion be allowed to get the proper alignment and therefore represents a more challenging case than that of the ferredoxins. The character of the conservation in these sequences is more variable than in the case of the ferredoxins making it a more interesting subject for profile analysis.

<Place Table II here>

Table II shows the ROC_{50} for the helicase family as a function of the size of the subset of sequences used to generate the profile. It is clear that both the average and evolutionary profile methods are able to extract a large amount of useful information from a fairly small set of sequences. Highly discriminatory profiles can be generated from as few as two to six sequences in the case of the evolutionary method. The evolutionary profile method is distinctly better than the average method for the helicase example; It is not until subsets of at least twelve sequences are

used that the average profiles equal the performance of the evolutionary profiles calculated from only two sequences.

Discussion

Comparison to Single Sequences

One can interpret the ROC_{50} statistic as the probability that a randomly selected positive sequence will score higher than a randomly selected negative sequence. Since the negative sequences are limited to only the highest scoring ones, the top 50 for the ROC_{50} , one could say that it is the probability that a truly homologous sequence will score higher than the most likely false positives. Table II shows that for database searches using single sequences as queries there is an average chance of about 20% that a homologous sequence will score below unrelated sequences, leading to a high chance of missing a homolog or misclassifying a sequence as related to a false positive. The high variability of matching to single sequences is also seen in the large standard deviation for the single sequence value. Adding information from as little as one or two additional sequences greatly improves the discriminatory power, and as importantly, greatly reduces the variability. These are important concerns in the development of motif descriptions suitable for the automatic annotation of sequences or homology based structural models.

Average Profiles vs. Evolutionary Profiles

The average profile method seeks to extract information from a single set of prior information embodied in the scoring table used in the averaging process. When the scoring table is based on the observed mutational exchanges between amino acid residues, as is the PAM 250 table typically used, it represents a superposition of all of the chemical similarities between the residues. A heuristic way to view the average method is that it seeks to discover, from all the superposed chemical similarities, the one property that is common at an aligned position. The average profile

achieves this because only residues that are chemical similar to each other will end up with high scores after the averaging process (see Gribskov [14] for examples).

Average profiles have been shown to be excellent discriminators for classifying protein families, usually achieving perfect or nearly perfect classification at unambiguous significance levels, for example Z scores of 7.5 and above. This classification ability is usually accompanied by a lower level of false positives than is found with regular expression methods (results not shown), and by a greater ability to detect distantly related sequences that may lack residues that, up to then, were absolutely conserved. These properties have led to the high level of interest in the further development of profile and profile-like models (e.g., Bairoch and Bucher [1]). The average profile method, however, clearly does not adequately emphasize positions that are highly conserved. Consider, for example, a residue that is absolutely conserved in every sequence in a family of 100 sequences. Such a position is required, often participating in critical structures or functions such as the active site of an enzyme. However, the average profile represents such a position with a row of values identical to the corresponding row for the conserved residue in the scoring table on which the profile is based (eq. 1). This inability to properly model highly conserved positions gave us the impetus to develop the evolutionary profile method.

The idea of the evolutionary profile method is to make a much more detailed and biologically relevant model of protein sequence families. There are two basic observations that guided the development of the evolutionary profile approach. Firstly, it is well known that the amount of conservation among protein sequences varies widely from position to position. Thus it can be said that the positions in a sequence evolve at different rates. Secondly, the type of conservation varies widely from position to position, i.e., there are different allowed residues at each position in a sequence, a constraint that arises primarily from the three dimensional structure.

The evolutionary profile method selects the set of matching residues at each position by fitting the observed distribution of residues to distributions predicted for all possible ancestral residues and PAM distances according to the Dayhoff evolutionary model. This generates a model of a sequence family in which each position can be interpreted in a biologically sensible and

intelligible way as a small set of preferred residues and evolutionary rates. A comparison of the alignment shown in figure 4 with the mixture components shown in figure 2 shows that the model closely corresponds to biological intuition. The highly conserved positions are modeled as mixtures of only one or two components at short evolutionary distances, while less conserved positions are modeled as mixtures of several components generally at longer evolutionary distances. It is noteworthy that evolutionary profiles can be easily scaled to longer or shorter evolutionary distances by simply multiplying or dividing the PAM distances fit during the modeling process. For instance, by simply multiplying all the fit PAM distances in figure 2 by a constant, and then recalculating the log-odds matrix, we can generate a model of the family at a greater evolutionary distance. We have not yet investigated this feature in detail, but it has the potential for allowing one to extend a model based on relatively closely related sequences to very distant members of a family.

Evolutionary profiles perform better than average profiles in generating discriminators for sequence classification (Table I and Table II). Clearly, using this approach, models with very good discriminatory power can be generated from very small numbers of sequences, a sharp contrast to the fairly large numbers of sequences required to train hidden Markov models. This ability to generalize from a small set of observed sequences is due to the incorporation of a strong biological model of sequence conservation. In the near future we will examine the possibility of incorporating other biologically relevant prior information such as known patterns of chemical similarity between the amino acid residues and predicted secondary structure within the same mixture model framework used for the evolutionary profile.

Comparison to Hidden Markov Models

The underlying model represented by a profile bears a close similarity to hidden Markov models (HMMs) recently introduced for describing protein families[4,5]. Each row in the profile can be regarded as a "match state", and the values in the row as the emission probabilities for each

of the twenty possible amino acid residues. The position specific gap weights represent transition probabilities for moving to an insert or delete state from a match state. The main difference between the profile model and the most common HMM is that the profile model requires that the transition from a match state to an insert state and the transition from a match state to a delete state have the same probability. Because an insertion in one sequence can be viewed as a deletion in another, hence the common term "indel", the profile model's requirement that the insert and delete transitions be equal seems reasonable (note, however, that the original profile model [12] did not make this requirement making it more similar to an HMM).

Profile Libraries

We are actively engaged in extending the available profile libraries. We currently have a library of over 600 protein motifs based on release 10 of PROSITE. These profiles were generated by locating the signature sequence for each of the PROSITE families in the annotated true positive sequences, extending these sequences by twenty residues on both sides of the signature, multiply aligning the sequences, and producing average profiles. Each of these profiles has been validated by database searches, and is available for use with PROFILESCAN. These profiles will be updated in the near future using the evolutionary profile method.

Acknowledgments

This work was supported by the National Science Foundation through cooperative agreement ASC-8902825 with the San Diego Supercomputer Center, and by NIH grant P41 RR08605. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author and do not necessarily reflect the views or policies of the National Science Foundation, the National Institutes of Health, or other supporters of the San Diego Supercomputer Center.

References

[1] A. Bairoch and P. Bucher, *Nucleic Acids Res.*, **22**, 3583-3589 (1994).

[2] S. Henikoff and J.G. Henikoff, *Proc.Natl.Acad.Sci. USA*, **89**, 10915-10919 (1992).

[3] E.L. Sonnhammer and D. Kahn, *Protein Science*, **3**, 482-492 (1994).

[4] A. Krogh, M. Brown, I.S. Mian, K. Sjolander and D. Haussler, *J.Molec.Biol.*, **235**, 1501-1531 (1994).

[5] P. Baldi ; Y. Chauvin ; T. Hunkapiller and M.A. McClure, *Proc.Natl.Acad.Sci. USA*, **91**, 1059-1063 (1994).

[6] J.U. Bowie, R. Luthy and D. Eisenberg, *Science*, **253**, 164-170 (1991).

[7] M. Wilmanns M, and D. Eisenberg, *Proc.Natl.Acad.Sci. USA*, **90**, 1379-1383 (1993).

[8] K.Y. Zhang and D. Eisenberg, *Protein Science*, **3**, 687-695 (1994).

[9] T.F. Smith and M.S. Waterman, *J.Molec.Biol*, **147**, 195-197 (1981).

[10] M. Gribskov, R. Lüthy and D. Eisenberg, this series, vol **183**, pp 146-159 (1990).

[11] The Profile Analysis package is available from the authors, although the programs are in the midst of conversion from FORTRAN to C programming languages. Please contact Michael Gribskov at gribskov@sdsc.edu for details on the status of implementation. Current program source code and libraries of profiles are available by FTP from ftp.sdsc.edu/pub/sdsc/biology.

The Profile Analysis package is also distributed by the Genetics Computer Group, Madison WI, as part of their sequence analysis package.

[12] M. Gribskov, A.D. McLachlan and D. Eisenberg, *Proc.Natl.Acad.Sci. USA*, **84,**4355-4358 (1987).

[13] M. Gribskov and D. Eisenberg, *in* "Techniques in Protein Chemistry" (T.E. Hugli, ed.), pp 108-117, Academic Press, San Diego (1989).

[14] M. Gribskov, *in* "Computer Analysis of Sequence Data, Part II" (A.M. Griffin and H.G. Griffin Eds), Methods in Molecular Biology Vol. 25,pp 247-266 (1994).

[15] PROFILE-SS is available from the Pittsburgh Supercomputing Center, contact Alex Ropelewski, ropelews@psc.edu for details, or using world wide web, access http://pscinfo.psc.edu/general/spftware/profiless/profiless.html.

[16] M. Gribskov, M. Homyak, J. Edenfield and D. Eisenberg, *CABIOS*, **4**, 61-66, 1988.

[17] M.O. Dayhoff, R.M. Schwartz, and B.C. Orcutt, *in* " Atlas of Protein Sequence and Structure", Vol. 5, Supp. 3, (M.O. Dayhoff, ed.), pp 345-358, National Biomedical Research Foundation, Washington DC (1978).

[18] J. Felsenstein, *Am.J.Human Genet.*, **25**, 471-492 (1973).

[19] M.H. Zweig and G. Campbell, *Clinical Chem.*, **39**, 561-577 (1993).

[20] M. Gribskov and N.L. Robinson, *Computers Chem.*, in Press (1995).

[21] M. Gribskov, *in* "Distance-Based Approaches to Protein Structure Determination II" (S. Brunak and H. Bohr, eds.), CRC Press, in Press (1995).

[22] H Beinert, *Faseb J.*, **4**, 2483-91 (1990).

Table I

Profile Method	Number of Sequences Included in Profile			
	3	6	12	134
Average	82.2 (5.6)	93.0 (2.0)	93.6 (1.5)	95.2
Evolutionary	84.2 (8.0)	95.2 (1.0)	95.6 (0.6)	98.3

Table I. Performance of average profiles and evolutionary profiles on the 4Fe-4S ferredoxin family (ROC_{50} x 100). Profiles were generated from aligned sequences selected at random from the 4Fe-4S ferredoxin family. Subsets of size 3, 6 and 12 as well as the entire family of sequences (134 members) were used to produce profiles. Searches of the Swiss-Prot database (release 31.0) using the program PROFILESEARCH were then performed with each of the profiles. The ability of the profile to identify sequences in the family was evaluated using ROC_{50} method. Values in the table are the ROC_{50} times 100 and represent mean and the standard deviation (in parentheses) of ten replicates of subsets of the indicated size.

Table II

Method	Single Sequence PAM 250	Number of Sequences Included in Profile				
		2	3	6	12	38
Average	78.0 (16.4)	86.6 (7.0)	91.2 (4.4)	95.6 (1.8)	97.4 (0.9)	97.7
Evolutionary	NA	97.2 (1.4)	98.2 (0.9)	99.2 (0.9)	99.3 (0.09)	99.3

Table II. Performance of average profiles and evolutionary profiles on the ATP dependent helicase family (ROC_{50} x 100). Profiles were generated from aligned sequences selected at random from the helicase family sequences shown in fig. 4. Subsets of size 2, 3, 6 and 12 as well as the entire family of sequences (38 members) were used to produce profiles. Searches of the Swiss-Prot database (release 31.0) using the program PROFILESEARCH were then performed with each of the profiles. The ability of the profile to identify sequences in the family was evaluated using ROC_{50} method. Values in the table are the ROC_{50} times 100 and represent mean and the standard deviation (in parentheses) of ten replicates of the indicated numbers of sequences. For comparison, the average ROC_{50} for all 38 sequences is shown.

Figure Legends

Figure 1. Evolutionary profile calculated for the sequences shown in figure 4. Each row corresponds to a column of the aligned sequence. The consensus sequence shown at the left represents the highest scoring column in each row and can be used as a cross-reference to figure 4. The most conserved regions of the sequence have the consensus sequence and the corresponding column shown in bold face.

Figure 2. Mixture density components for the evolutionary profile of the helicases DEAD region. This figure corresponds to the profile shown in figure 1 and the alignment shown in figure 4. Each line shows in rank order, the components of the mixture model at that position. Components are given as A D (W), where A is the ancestral residues, D is the fit PAM distance, and W is the weight of the component in the mixture distribution. Note that the component of the mixture with the highest weight does not necessarily correspond to the highest scoring column in the profile.

Figure 3. ROC plot for two ferredoxin profiles. The solid line shows the curve of a profile calculated by the average method using only three sequences ($ROC_{50} = 0.71$), an example of a relatively poor discrimination. The dashed line shows the ROC plot for an evolutionary profile based on all 134 ferredoxin sequences in the database ($ROC_{50} = 0.99$), an example of nearly perfect discrimination.

Figure 4. Alignment of helicases in the region of the conserved "DEAD" sequence. The most conserved regions are highlighted in bold face. The bottom row, labeled consensus, represents the highest scoring column in the evolutionary profile shown in figure 1.

Figure 1

ns	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Gap	Len
G	9	-38	4	-2	-44	52	-22	-28	-15	-36	-28	3	-6	-14	-23	8	-2	-18	-53	-45	100	100
P	10	19	0	-1	-35	3	-8	-6	-10	-16	-10	0	30	-3	-9	8	7	0	-46	-26	100	100
H	-15	-60	39	37	-56	-16	53	-37	-4	-44	-32	26	-13	42	-3	-8	-16	-36	-69	-32	100	100
I	-21	-54	-53	-44	-8	-50	-53	91	-50	26	19	-43	-42	-43	-45	-41	-15	46	-94	-28	100	100
V	-15	-55	-56	-46	-19	-49	-53	60	-53	28	20	-47	-43	-43	-48	-44	-15	72	-105	-36	100	100
V	-16	-55	-63	-51	-30	-54	-65	69	-62	16	14	-53	-51	-54	-56	-49	-15	78	-125	-49	100	100
A	66	-68	-16	-18	-75	37	-61	-56	-52	-71	-57	-17	-14	-42	-63	6	-3	-29	-100	-80	100	100
T	-50	-106	-84	-88	-130	-75	-109	-97	-98	-120	-101	-75	-74	-99	-105	-50	1**17**	-89	-46	-126	100	100
P	-16	-81	-61	-55	-110	-53	-62	-83	-72	-83	-85	-56	1**15**	-46	-61	-26	-40	-62	-117	-104	100	100
G	-99	-15	-104	-110	-151	1**02**	-5	-9	-125	-141	-10	-107	-111	-121	-8	-101	-110	-123	-38	-26	100	100
R	-98	-144	-7	-10	-141	-105	-68	-98	0	-111	-60	-56	-91	-59	1**29**	-75	-74	-110	-84	-69	100	100
L	-24	-70	-62	-48	5	-56	-47	34	-51	69	39	-47	-39	-30	-46	-45	-24	32	-94	-21	100	100
L	-25	-65	-55	-44	8	-51	-42	30	-46	70	41	-42	-35	-26	-41	-41	-23	20	-86	-17	100	100
D	-42	-94	105	33	-94	-34	-19	-63	-33	-78	-68	14	-48	0	-52	-32	-41	-66	-106	-74	100	100
L	-17	-38	-28	-25	35	-32	-9	16	-26	47	32	-21	-20	-14	-22	-23	-15	10	-30	33	100	100
L	-23	-61	-51	-41	1	-48	-41	33	-44	69	38	-40	-33	-26	-40	-38	-21	22	-93	-26	100	100
Q	-8	-52	27	35	-53	-11	16	-25	17	-32	-12	12	-7	37	14	-4	-7	-26	-46	-40	100	100
K	-26	-62	-8	-9	-65	-32	1	-29	59	-41	2	8	-18	7	51	-8	-8	-38	-34	-62	100	100
G	8	-26	13	5	-38	21	-4	-18	-3	-28	-17	16	0	-3	-9	17	9	-14	-41	-32	100	100
T	4	-15	6	2	-25	1	0	-4	6	-10	-1	6	2	0	3	7	11	-3	-23	-19	100	100
V	-3	-37	33	21	-40	-7	-11	20	-15	-7	-1	8	-17	0	-25	-9	-3	36	-74	-36	22	22
T	11	-29	10	6	-39	16	-13	-7	-7	-24	-12	7	0	-4	-18	12	25	-3	-57	-35	22	22
K	-21	-61	-10	-13	-50	-32	-14	-18	60	-21	11	8	-21	0	32	-6	3	-26	-48	-53	22	22
G	9	-25	4	0	-29	26	-16	2	-13	-9	-4	1	0	-8	-19	5	4	8	-56	-29	22	22
L	-10	-37	-28	-24	16	-27	-23	27	-25	47	33	-22	-18	-17	-23	-20	-8	25	-52	3	22	22
K	-8	-36	10	3	-40	-9	7	-15	32	-21	0	19	-6	7	26	1	0	-18	-26	-33	100	100
L	-24	-63	-52	-42	9	-49	-39	23	-44	71	41	-41	-33	-25	-40	-39	-24	17	-79	-13	100	100
K	-5	-38	18	13	-44	-4	7	-17	25	-25	-4	17	-4	12	20	2	0	-19	-32	-36	100	100
K	-10	-40	5	3	-42	-13	16	-17	34	-23	1	10	-4	13	34	-1	-1	-20	-19	-36	100	100
V	-6	-9	-38	-32	-14	-32	-42	52	-39	27	23	-32	-29	-33	-36	-25	-4	58	-89	-26	100	100
K	-20	-58	4	0	-62	-23	1	-27	56	-38	1	15	-16	8	33	-4	-4	-34	-40	-56	100	100
L	-17	-38	-34	-30	40	-35	-20	21	-29	46	36	-25	-24	-21	-27	-25	-15	15	-27	40	100	100
L	-27	-70	-58	-46	11	-54	-42	27	-48	72	41	-44	-36	-27	-43	-43	-26	17	-80	-12	100	100
V	-44	-89	-95	-83	-61	-85	-93	38	-94	-11	-13	-86	-82	-84	-86	-84	-45	103	-150	-77	100	100
L	-48	-105	-92	-71	-9	-82	-65	8	-74	89	40	-69	-57	-41	-67	-70	-50	0	-126	-46	100	100
D	-76	-59	1**24**	-13	-60	-74	-61	-98	-72	-42	-34	-19	-84	-43	-21	-69	-78	-101	-72	-39	100	100
E	-83	-57	-14	1**22**	-60	-79	-62	-106	-77	-116	-29	-48	-89	-34	-12	-79	-86	-108	-69	-120	100	100
A	9**9**	-111	-76	-75	-128	-53	-107	-102	-99	-119	-101	-74	-63	-94	-106	-45	-45	-80	-141	-125	100	100
D	-76	-59	1**24**	-13	-60	-74	-61	-98	-72	-42	-34	-19	-84	-43	-21	-69	-78	-101	-72	-39	100	100
R	-40	-76	-23	-24	-79	-49	-3	-38	56	-55	-4	-1	-26	2	82	-13	-18	-51	-28	-85	100	100
M	-30	-79	-63	-50	2	-59	-47	24	-47	72	78	-49	-39	-28	-45	-47	-29	15	-103	-28	100	100
L	-47	-100	-90	-70	-5	-80	-63	8	-73	88	34	-67	-56	-40	-66	-69	-50	0	-110	-37	100	100
D	-14	-59	71	46	-66	-8	-2	-37	-10	-52	-41	30	-22	12	-28	-2	-12	-39	-73	-49	100	100
L	-12	-46	-31	-26	5	-31	-26	26	-22	52	51	-25	-20	-16	-23	-22	-10	23	-71	-12	100	100
G	1	-73	-8	-17	-68	79	-55	-61	-45	-68	-64	-9	-30	-39	-62	0	-26	-40	-97	-81	100	100
F	-63	-76	-81	-77	120	-81	-54	-19	-77	8	-1	-65	-68	-67	-70	-65	-62	-37	-50	26	100	100
G	-1	-20	10	9	-24	0	9	-8	10	-8	0	7	1	12	11	0	0	-7	-17	-18	100	100
Q	-5	-45	47	44	-51	-1	5	-25	-2	-36	-26	24	-10	19	-14	-1	-7	-26	-57	-37	100	100
E	-9	-62	47	50	-62	-10	16	-33	-3	-41	-29	18	-13	42	-10	-8	-14	-33	-68	-43	100	100
I	-15	-50	-43	-35	2	-41	-40	64	-38	41	32	-34	-31	-30	-35	-32	-11	44	-82	-17	100	100
D	-1	-39	29	27	-39	1	12	-17	5	-24	-13	15	-2	19	0	1	-1	-17	-43	-27	100	100
Q	-9	-47	18	18	-48	-11	22	-22	24	-29	-7	17	-6	31	22	-1	-4	-24	-35	-36	100	100
I	-20	-53	-55	-45	-19	-50	-58	94	-52	21	17	-44	-44	-46	-47	-42	-14	49	-109	-41	100	100
L	-13	-47	-38	-31	11	-37	-32	42	-32	46	37	-30	-26	-24	-30	-27	-10	40	-66	-5	100	100
K	-3	-28	4	1	-37	-4	1	-12	29	-19	1	8	-1	5	25	3	4	-13	-23	-33	100	100
L	-4	-18	-4	-4	1	-9	3	10	-6	23	17	-3	-4	-1	-4	-6	-2	9	-28	1	100	100
L	-24	-70	-62	-48	3	-55	-47	34	-51	69	40	-48	-39	-30	-46	-45	-23	33	-98	-23	100	100
*	16	2	34	18	15	21	7	31	17	44	12	4	9	8	21	8	14	24	0	3		

Figure 2

G 128 (0.53)	S 256 (0.12)	A 256 (0.11)	K 512 (0.08)				
P 256 (0.34)	A 256 (0.31)	S 256 (0.13)	C 512 (0.12)	V 256 (0.08)			
H 128 (0.21)	Q 128 (0.19)	D 128 (0.18)	E 128 (0.17)	N 128 (0.12)	K 512 (0.07)		
I 32 (0.36)	V 128 (0.32)	L 256 (0.27)					
V 64 (0.51)	L 256 (0.32)	I 64 (0.16)					
V 64 (0.59)	I 64 (0.23)	L 256 (0.16)					
A 32 (0.51)	G 128 (0.36)	S 128 (0.07)					
T 1 (0.86)	A 128 (0.06)	S 128 (0.06)					
P 16 (0.81)	A 128 (0.10)						
G 1 (0.94)							
R 1 (0.79)	K 128 (0.18)						
L 128 (0.69)	V 128 (0.18)	I 128 (0.08)					
L 128 (0.72)	V 256 (0.15)	I 128 (0.08)					
D 16 (0.59)	E 128 (0.19)	N 128 (0.07)					
L 256 (0.59)	F 256 (0.17)	Y 256 (0.07)	V 512 (0.07)				
L 128 (0.69)	V 256 (0.18)	I 128 (0.10)					
K 256 (0.27)	E 128 (0.21)	Q 128 (0.16)	D 256 (0.13)	N 256 (0.06)	H 256 (0.05)		
K 128 (0.63)	R 128 (0.17)	Q 256 (0.05)					
G 256 (0.28)	S 128 (0.18)	K 512 (0.10)	A 256 (0.10)	N 128 (0.09)	T 256 (0.07)	D 256 (0.06)	
T 256 (0.33)	K 512 (0.31)	S 256 (0.09)	A 512 (0.08)				
V 128 (0.34)	D 128 (0.24)	E 256 (0.17)	I 256 (0.07)	A 512 (0.06)			
A 256 (0.23)	T 128 (0.22)	G 256 (0.22)	S 256 (0.09)	D 256 (0.08)	E 256 (0.05)		
K 128 (0.71)	L 512 (0.11)	T 128 (0.06)	R 256 (0.06)				
G 256 (0.40)	A 256 (0.25)	V 256 (0.20)					
L 256 (0.57)	V 256 (0.22)	I 256 (0.10)	F 256 (0.07)				
K 256 (0.54)	N 128 (0.13)	R 256 (0.08)	D 512 (0.05)				
L 128 (0.72)	V 256 (0.14)	I 256 (0.06)					
K 256 (0.39)	D 256 (0.11)	E 256 (0.10)	N 128 (0.07)	R 256 (0.07)	S 256 (0.06)	G 512 (0.06)	Q 256 (0.06)
K 256 (0.55)	R 256 (0.18)	H 256 (0.07)	N 256 (0.06)	Q 256 (0.06)			
V 128 (0.55)	L 256 (0.20)	I 128 (0.18)					
K 128 (0.59)	R 256 (0.11)	E 256 (0.05)	Q 256 (0.05)				
L 256 (0.52)	F 256 (0.22)	Y 256 (0.10)	V 256 (0.08)	I 256 (0.06)			
L 128 (0.76)	V 256 (0.11)	F 256 (0.05)	I 128 (0.05)				
V 16 (0.78)	L 256 (0.11)	I 64 (0.10)					
L 64 (0.88)	V 256 (0.06)						
D 1 (0.80)	E 128 (0.10)						
E 1 (0.83)	D 128 (0.09)						
A 4 (0.88)							
D 1 (0.80)	E 128 (0.10)						
K 128 (0.55)	R 64 (0.37)						
L 128 (0.76)	V 256 (0.10)	M 32 (0.08)					
L 64 (0.87)	V 256 (0.06)						
D 64 (0.37)	E 128 (0.23)	N 128 (0.11)	G 512 (0.07)	S 128 (0.06)			
L 256 (0.69)	V 256 (0.15)	M 128 (0.08)	I 256 (0.07)				
G 64 (0.78)	A 256 (0.09)	S 256 (0.05)					
F 16 (0.68)	L 256 (0.20)	Y 256 (0.09)					
K 512 (0.46)	Q 256 (0.12)	E 512 (0.11)	R 512 (0.08)	D 512 (0.07)			
D 128 (0.25)	E 128 (0.25)	G 512 (0.11)	N 128 (0.08)	K 512 (0.07)	Q 256 (0.07)	A 512 (0.05)	
E 128 (0.33)	D 128 (0.22)	Q 128 (0.19)	K 512 (0.08)	N 256 (0.07)	H 256 (0.05)		
L 256 (0.48)	V 128 (0.27)	I 64 (0.22)					
E 256 (0.29)	D 256 (0.23)	K 512 (0.15)	N 256 (0.12)	Q 256 (0.11)	H 512 (0.06)		
K 256 (0.34)	Q 128 (0.13)	E 256 (0.12)	D 256 (0.11)	R 256 (0.09)	H 256 (0.09)	N 128 (0.07)	
I 32 (0.39)	V 128 (0.38)	L 256 (0.21)					
L 256 (0.56)	V 128 (0.21)	I 128 (0.14)	F 256 (0.05)				
K 256 (0.47)	R 256 (0.09)	S 256 (0.09)	T 256 (0.09)	G 512 (0.08)	A 512 (0.05)		
L 512 (0.62)	V 512 (0.14)	I 512 (0.07)	H 256 (0.06)				
L 128 (0.69)	V 128 (0.18)	I 128 (0.08)					

Figure 3

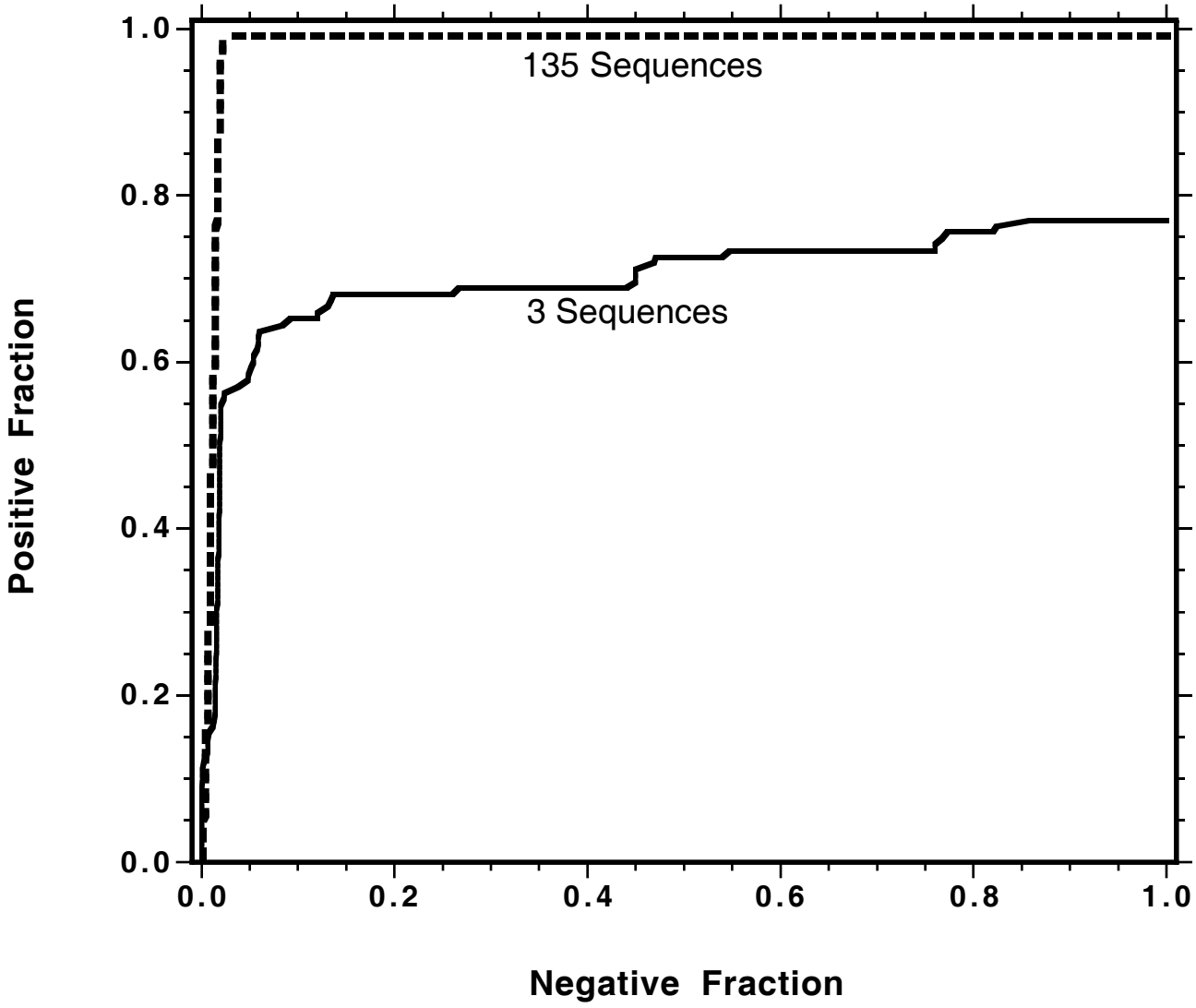
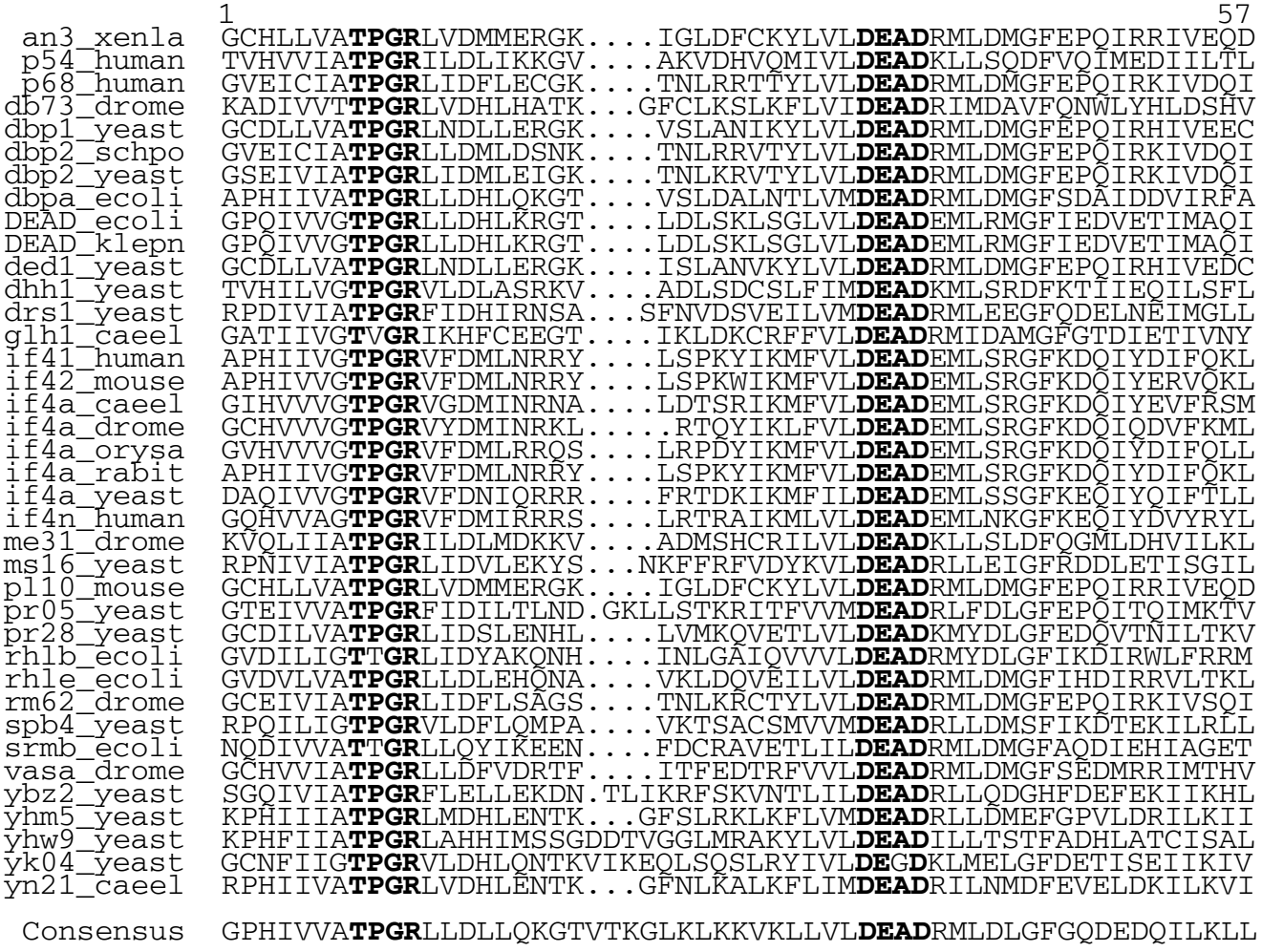

1.0
0.8
0.6
0.4
0.2
0.0
Positive Fraction
135 Sequences
3 Sequences
0.0
0.2
0.4
0.6
0.8
1.0
Negative Fraction

Figure 4


1
57
an3_xenla GCHLLVATPGRLVDMMERGK....IGLDFCKYLVLDEADRMLDMGFEPQIRRIVEQD
p54_human TVHVVIATPGRILDLIKKGV....AKVDHVQMIVLDEADKLLSQDFVQIMEDIILTL
p68_human GVEICIATPGRLIDFLECGK....TNLRRTTYLVLDEADRMLDMGFEPQIRKIVDQI
db73_drome KADIVVTTPGRLVDHLHATK...GFCLKSLKFLVIDEADRIMDAVFQNWLYHLDSHV
dbp1_yeast GCDLLVATPGRLNDLLERGK....VSLANIKYLVLDEADRMLDMGFEPQIRHIVEEC
dbp2_schpo GVEICIATPGRLLDMLDSNK....TNLRRVTYLVLDEADRMLDMGFEPQIRKIVDQI
dbp2_yeast GSEIVIATPGRLIDMLEIGK....TNLKRVTYLVLDEADRMLDMGFEPQIRKIVDQI
dbpa_ecoli APHIIVATPGRLLDHLQKGT....VSLDALNTLVMDEADRMLDMGFSDAIDDVIRFA
DEAD_ecoli GPQIVVGTPGRLLDHLKRGT....LDLSKLSGLVLDEADEMLRMGFIEDVETIMAQI
DEAD_klepn GPQIVVGTPGRLLDHLKRGT....LDLSKLSGLVLDEADEMLRMGFIEDVETIMAQI
ded1_yeast GCDLLVATPGRLNDLLERGK....ISLANVKYLVLDEADRMLDMGFEPQIRHIVEDC
dhh1_yeast TVHILVGTPGRVLDLASRKV....ADLSDCSLFIMDEADKMLSRDFKTIIEQILSFL
drs1_yeast RPDIVIATPGRFIDHIRNSA...SFNVDSVEILVMDEADRMLEEGFQDELNEIMGLL
glh1_caeel GATIIVGTVGRIKHFCEEGT....IKLDKCRFFVLDEADRMIDAMGFGTDIETIVNY
if41_human APHIIVGTPGRVFDMLNRRY....LSPKYIKMFVLDEADEMLSRGFKDQIYDIFQKL
if42_mouse APHIVVGTPGRVFDMLNRRY....LSPKWIKMFVLDEADEMLSRGFKDQIYERVQKL
if4a_caeel GIHVVVGTPGRVGDMINRNA....LDTSRIKMFVLDEADEMLSRGFKDQIYEVFRSM
if4a_drome GCHVVVGTPGRVYDMINRKL.....RTQYIKLFVLDEADEMLSRGFKDQIQDVFKML
if4a_orysa GVHVVVGTPGRVFDMLRRQS....LRPDYIKMFVLDEADEMLSRGFKDQIYDIFQLL
if4a_rabit APHIIVGTPGRVFDMLNRRY....LSPKYIKMFVLDEADEMLSRGFKDQIYDIFQKL
if4a_yeast DAQIVVGTPGRVFDNIQRRR....FRTDKIKMFILDEADEMLSSGFKEQIYQIFTLL
if4n_human GQHVVAGTPGRVFDMIRRRS....LRTRAIKMLVLDEADEMLNKGFKEQIYDVYRYL
me31_drome KVQLIIATPGRILDLMDKKV....ADMSHCRILVLDEADKLLSLDFQGMLDHVILKL
ms16_yeast RPNIVIATPGRLIDVLEKYS...NKFFRFVDYKVLDEADRLLEIGFRDDLETISGIL
pl10_mouse GCHLLVATPGRLVDMMERGK....IGLDFCKYLVLDEADRMLDMGFEPQIRRIVEQD
pr05_yeast GTEIVVATPGRFIDILTLND.GKLLSTKRITFVVMDEADRLFDLGFEPQITQIMKTV
pr28_yeast GCDILVATPGRLIDSLENHL....LVMKQVETLVLDEADKMYDLGFEDQVTNILTKV
rhlb_ecoli GVDILIGTTGRLIDYAKQNH....INLGAIQVVVLDEADRMYDLGFIKDIRWLFRRM
rhle_ecoli GVDVLVATPGRLLDLEHQNA....VKLDQVEILVLDEADRMLDMGFIHDIRRVLTKL
rm62_drome GCEIVIATPGRLIDFLSAGS....TNLKRCTYLVLDEADRMLDMGFEPQIRKIVSQI
spb4_yeast RPQILIGTPGRVLDFLQMPA....VKTSACSMVVMDEADRLLDMSFIKDTEKILRLL
srmb_ecoli NQDIVVATTGRLLQYIKEEN....FDCRAVETLILDEADRMLDMGFAQDIEHIAGET
vasa_drome GCHVVIATPGRLLDFVDRTF....ITFEDTRFVVLDEADRMLDMGFSEDMRRIMTHV
ybz2_yeast SGQIVIATPGRFLELLEKDN.TLIKRFSKVNTLILDEADRLLQDGHFDEFEKIIKHL
yhm5_yeast KPHIIIATPGRLMDHLENTK...GFSLRKLKFLVMDEADRLLDMEFGPVLDRILKII
yhw9_yeast KPHFIIATPGRLAHHIMSSGDDTVGGLMRAKYLVLDEADILLTSTFADHLATCISAL
yk04_yeast GCNFIIGTPGRVLDHLQNTKVIKEQLSQSLRYIVLDEGDKLMELGFDETISEIIKIV
yn21_caeel RPHIIVATPGRLVDHLENTK...GFNLKALKFLIMDEADRILNMDFEVELDKILKVI
Consensus GPHIVVATPGRLLDLLQKGTVTKGLKLKKVKLLVLDEADRMLDLGFGQDEDQILKLL