Identification of Protein Motifs Using Conserved Amino Acid Properties and Partitioning Techniques

Thomas D. Wu

Section on Medical Informatics, Stanford University
Medical School Office Building, Room X-215
Stanford, California 94305
twu@camis.stanford.edu

Douglas L. Brutlag

Department of Biochemistry, Stanford University
Beckman Center, Room B403
Stanford, California 94305
brutlag@cmgm.stanford.edu

Abstract

Analyzing a set of protein sequences involves a fundamental relationship between the coherency of the set and the specificity of the motif that describes it. Motifs may be obscured by training sets that contain incoherent sequences, in part due to protein subclasses, contamination, or errors. We develop an algorithm for motif identification that systematically explores possible patterns of coherency within a set of protein sequences. Our algorithm constructs alternative partitions of the training set data, where one subset of each partition is presumed to contain coherent data and is used for forming a motif. The motif is represented by multiple overlapping amino acid groups based on evolutionary, biochemical, or physical properties. We demonstrate our method on a training set of reverse transcriptases that contains subclasses, sequence errors, misalignments, and contaminating sequences. Despite these complications, our program identifies a novel motif for the subclass of retroviral and retrovirus-related reverse transcriptases. This motif has a much higher specificity than previously reported motifs and suggests the importance of conserved hydrophilic and hydrophobic residues in the structure of reverse transcriptases.

Introduction

Finding patterns, or motifs, in protein sequences involves two essential steps: assembling a training set of sequences with common structure or function and then analyzing the training set for regions of conserved amino acid residues. Hence, the resulting motif depends critically on the training set used. In fact, there is a fundamental relationship between the coherency of a training set and the specificity of a motif. When a training set contains incoherent sequences, a motif must become less specific in order to describe the entire set. A training set should ideally contain a representative sample from a coherent class of proteins, but obtaining a coherent set is complicated by several characteristics of protein sequence data:

1. Protein classes may contain subclasses. Each subclass may have a specific motif, whereas the entire class may have no motif or only a very general one.

2. The training set may be contaminated. Structural or functional evidence for including a sequence in a training set is often imprecise, so some sequences in the training set may not belong with the others. A contaminating sequence may mask a specific motif present in the other sequences.

3. Sequence data may contain errors. Errors may arise by misaligning a sequence, or by substituting, inserting, or deleting an amino acid residue. Such errors may obscure an underlying specific motif.

Because of these complications in the data, researchers have developed compensatory methods for analyzing training sets. One way to handle incoherent data is to resort to probabilistic motifs [Henikoff & Henikoff 1991] or profiles [Gribskov, Luthy, & Eisenberg 1990]. Profiles have been used to generate training sets and identify motifs simultaneously [Tatusov, Altschul, & Koonin 1994]. However, probabilistic representations give poor insight into the structure or function of a protein. Moreover, probabilistic methods do not recognize or reject incoherent data, but rather incorporate them into their motif. Thus, subclasses, contamination, and errors still lower the accuracy and precision of probabilistic representations.

The alternative to probabilistic representations, discrete *motifs*, use categorical amino acid groups to describe sequences. Each amino acid group is a disjunctive set of amino acids permitted at a given position. However, the choice of amino acid groups is problematic. When amino acid groups are arbitrary, as allowed by the consensus sequence method [Bairoch 1991], the space of possibilities is enormous: For 20 amino acids there are over a million possible groups. Choosing among these possibilities with statistical validity requires extremely large training sets. But in practice, training sets usually contain fewer than 100 sequences, so motifs generated this way are highly underdetermined. On the other hand, some methods restrict the allowed collection of amino acid groups too much. For example, the motif identification method of Smith RF and Smith TF [1990] allows only one alphabet of groups, which do not overlap: `[DE]`, `[KRH]`, `[NQ]`, `[ST]`, `[VLI]`, `[FYW]`, `[AG]`, `[P]`, `[M]`, and `[C]`. But this highly limited collection of amino acid groups may miss important patterns, such as the set of small hydrophobic amino acids: `[VLIM]`. Another discrete method [Saqi & Sternberg 1994], which finds protein subclasses through cluster

This work was supported in part by grants LM 05716 and LM 07033 *from the National Library of Medicine*

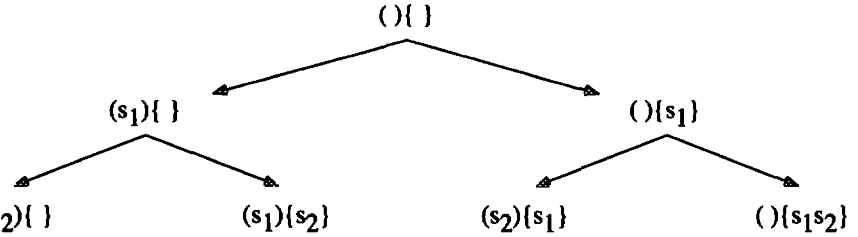

Figure 1 Search tree for partitioning sequence data. Each node in the tree contains two sets. The first set, denoted by ( ), is the inclusion set. The second set, denoted by { }, is the exclusion set. Each sequence is processed sequentially and placed in either the inclusion set or exclusion set.

analysis of subsequences in the training set, is also limited by its use of non-overlapping amino acid groups. In addition, that method cannot take advantage of information contained in a multiple sequence alignment. Existing discrete methods are generally weak because they employ amino acid groups that are not well founded on the range of known biochemical properties. One exception is work by Taylor [1986], who does use multiple overlapping amino acid groups based on known properties. However, his groups are based on a single Venn diagram of amino acid relationships. Furthermore, the amino acid groups at each position must cover the entire training set, so an incoherent sequence can easily make his motifs relatively nonspecific.

In this paper, we show how principled amino acid groups can be used to handle the issues of subclasses, contamination, and erroneous data. We contend that regions are conserved precisely because they maintain some biochemical or physical constraints required for the structure or function of a protein. Thus, we require that amino acid groups be based on some evolutionary, biochemical, or physical property, such as volume, charge, or hydrophobicity. Using these groups, our method systematically explores possible patterns of coherency within a training set. It constructs incrementally alternative partitions of the training set data into two subsets. One subset is presumed to contain coherent data and is used to construct a motif, whereas the other subset is presumed to contain incoherent data and disregarded. The partitions are compared with one another to find those that include as many sequences in the training set as possible while generating a motif that is relatively specific. Our work builds upon previous work on partitioning methods in automated problem solving [Wu 1990].

We have implemented our method as a computer program called SEQCLASS. In this paper, we describe our method and present the results of an example taken from the class of proteins called reverse transcriptases. This is a heterogenous class of proteins that contain at least three known subclasses: retroviral reverse transcriptases, class I retrotransposons, and class II retrotransposons. Our program is able to distinguish the first two subclasses from the third from the training set data without supervision. The resulting motif is highly sensitive and specific for this set of retroviral and retrovirus-related reverse transcriptases and gives insight into their structure.

The Algorithm

Training Set Data. Our method requires as input an aligned set of sequences, taken from a class of proteins with related structure or function. This training set may be aligned by some biological measure, such as a binding site or reactive center; by a probabilistic method, such as GENALIGN; or by a preliminary method for identifying discrete motifs, such as the method of [Smith HO, Annau, & Chandrasegaran 1990].

Amino Acid Group Database. Our method also requires a database of amino acid groups. Each group represents a set of amino acids that are closely related by some functional or evolutionary criterion. For instance, `[VLIM]` represents the small hydrophobic amino acids. Amino acids may belong to more than one amino acid group, and amino acid groups may overlap or subsume one another. Thus, the amino acids can be grouped by multiple criteria, such as charge, size, and hydrophobicity. Also, our database includes a set of twenty singleton amino acid groups, such as `[V]`, each containing only one of the twenty amino acids. Singleton groups enable the program to detect exact conservation of an amino acid at a position.

Search Space. Our method explores the space of possible ways to partition the training set data into two sets. The **inclusion set** contains sequences used to form a possible motif, while the **exclusion set** contains sequences presumed to be incoherent and therefore not used to form the motif. The method processes each sequence in the training set sequentially and constructs several partitions incrementally. Given a new sequence, the algorithm can modify each incremental partition in two possible ways: placing the sequence in the inclusion set or placing it in the exclusion set. This framework for generating partitions incrementally can be considered as a search tree, as shown in Figure 1, with each node in the tree representing an incremental partition. Level *n* in the tree contains incremental partitions of the first *n* sequences considered. Each node at that level has an inclusion and exclusion set that partition the first *n* sequences in a different way.

Motif Generation. Each node represents a possible motif, which is a consensus of the sequences in its inclusion set. To form the motif, each position is considered independently. The set of residues occurring at each position is compared against the database of allowed

```
Included sequences [4]:
APOL            P I R Q A F P Q C T I L Q Y M D D I L L A S P S H E D L L L
TLPH            P M R K M F P T S T I V Q Y M D D I L L A S P T N E E L Q Q
POBO            Q V S A A F S Q S L L V S Y M D D I L I A S P T E E Q R S Q
SADP            K V R H A W K Q M Y I I H Y M D D I L I A G K D G Q Q V L Q

Motif:                            1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3
                1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0   Chi-square  Elements
Identity:                                 Y M D D I L   A                     1.6069E+5
Group c:                                                  X                   8.2714E+0   PAGST
Group e:                                                            X X       3.2010E+2   QNED
Group g:                            X X               X                       3.9114E+2   VLI
Group h:          X                                                           5.0293E+2   VLIM
Group j:                  X                                                   2.5502E+3   FYW
Hydrophil:      X   X X     X X         X                   X X X             1.5425E+1   PAGSTQNEDHKR
------------------------------------------------------------------------------------------------------
Significance(Motif): 49.462
```

Figure 2 Example of motif generation. The four sequences in the inclusion set are aligned at the top. The motif is generated by comparing the residues at each position to the allowed amino acid groups. The most specific group or groups that apply at each consensus position are denoted by an X, unless a singleton group applies, in which case the corresponding letter is shown. The chi-squared value for a group compares the number of X's observed to the number expected, assuming a random set of sequences. This value indicates the statistical significance for that particular amino acid group. The overall significance of the motif is also computed as discussed in the text.

amino acid groups. If the residues all fall within an amino acid group, that group is said to **apply** to that position in the motif, and the position is said to be a **consensus position**. For example, if all residues in a given position are either valine or leucine, then the amino acid group [VLI] applies. Since amino acid groups may overlap and subsume one another, more than one amino acid group may apply, for example, the groups [VLI], [VLIM], and [CVLIMFYW]. In that case, we identify the most specific amino acid group, that which is subsumed by all others (i.e., [VLI]). Note that there may be more than one specific group. If [CVL] were present in our database, both it and [VLI] would be most specific. We show an example of motif generation in Figure 2.

Generalization and Pruning. When the program adds a sequence to an inclusion set of an incremental partition, it checks whether the sequence already matches the motif for that partition. If so, the program neither needs to revise the motif nor needs to generate the alternative partition that places the sequence in the exclusion set.

On the other hand, when a sequence is added to an inclusion set and the motif is revised, the program attempts to transfer sequences from the exclusion set to the inclusion set, in a process called **generalization**. Generalization is possible because the revised motif will necessarily be more general than before. The revised motif may then match some sequences in the exclusion set that it previously could not. Thus, after the program revises a new motif for a given node, it checks each sequence in the exclusion set to see if it matches the new motif. If so, it transfers the sequence from the exclusion set to the inclusion set.

Because of this generalization step, the same partition may be generated by more than one node. The program therefore checks partitions at each level for duplicates and removes, or prunes, them from the search tree. Pruning also occurs when a revised motif no longer contains any consensus positions. This situation signifies that the sequences in the corresponding inclusion set are mutually incoherent and that no motif can be obtained by modifying that partition any further.

Search Strategy. The space of possible partitions grows as an exponential function of the size of the training set. Exponentially large search spaces cannot be explored completely for arbitrarily large training sets. Thus, only part of the theoretical search space can be investigated. There are several different strategies to explore large search spaces. We currently use a strategy called beam search. This strategy sets an upper limit, called the beam width *w*, on the number of partitions at each level of the search tree. The best *w* partitions at each level are used for generating the next level of partitions.

Sequence Ordering. Search strategies can be affected by the order in which data is presented to them. If the initial data are skewed, the search algorithm may be biased towards finding a locally optimal solution rather than a globally optimal one. To help avoid this problem, we order the sequences based on their homology to the other sequences in the training set. The homology between two sequences can be measured by summing the accepted point mutation (PAM) values [Jones, Taylor, & Thornton 1992] over each pair of aligned amino acids. The search algorithm begins with the "central" sequence, the one with the greatest average homology to the other sequences in the training set. At each successive step, the algorithm then processes the "next closest" sequence, the one with the greatest average homology to the sequences already processed.

Evaluation Function. A given partition of the sequence data may be better than others, perhaps because it includes more sequences from the training set or because its motif is more specific. To determine which partitions are best, we score each node numerically. The evaluation function we currently use measures the goodness of a partition by computing the probability that its motif would have occurred in all sequences in the inclusion set. Let *s* be the number of sequences in the inclusion set and let motif M contain elements a_i, where a_i is the most specific amino acid group that applies at position i. If more than one specific amino acid group applies, then the one with the lowest probability is used. We compute the probability of each amino acid group as the sum of the frequencies of its constituent amino acids. The probability of the motif is then the product of probabilities for the amino acid group at each position:

$$p(M) = \prod_{a_i \text{ in } M} \left[\sum_{AA \text{ in } a_i} p(AA) \right]$$

where AA is an amino acid in the most specific group a_i and its frequency $p(AA)$ is taken from some data source (We currently use the frequencies of amino acids over the entire SWISSPROT 30 protein database). Hence the probability of the motif occurring in s sequences is $p(M)^{(s-1)}$. Since the probabilities are small and can overrun the floating point capabilities of some computers, we use the logarithm of the probabilities instead. We call this quantity the **significance** of the motif:

$$\text{sig}(M) = -(s-1)\log p(M)$$

Higher significance values indicate better motifs. Our function therefore rewards motifs that identify several consensus positions, contain specific amino acid groups, or include many sequences. Of course, other evaluation functions could also be used with our method, perhaps ones based on heuristic methods or information theory.

Implementation. We have implemented the above algorithm in the Common Lisp programming language as a program called SEQCLASS. To improve efficiency, we represent the set of amino acid groups as a bit vector. For each of the twenty amino acids, our program precomputes all groups that the amino acid belongs to and stores the resulting bit vector. We represent a motif as a list of consensus positions and corresponding amino acid group bit vectors. To revise a motif with a new sequence, our program then takes the intersection of the group bit vectors for each amino acid residue in the sequence and for the corresponding consensus position in the motif. We also represent inclusion and exclusion sets as bit vectors, so that duplicate partitions can be detected rapidly.

Experiment: Reverse Transcriptases

In order to test our program and demonstrate its capabilities, we use an example taken from a previously published method for motif identification [Smith HO, Annau, & Chandrasegaran 1990]. This example contains a training set of 33 reverse transcriptases, reproduced in Table 1. Reverse transcriptases are enzymes that create DNA polymers from RNA templates and are hence also known as RNA-directed DNA polymerases. Subclasses of reverse transcriptases have been identified in molecular biology. One subclass derives from retroviruses. Another subclass, called class I retrotransposons, are found in eukaryotic cells, including yeast, Drosophila, plants, and mammals. These enzymes are also known as retrovirus-related reverse transcriptases. Both retroviral and retrovirus-related reverse transcriptases have terminal repeat sequences at both ends. A third subclass, called class II retrotransposons, lacks these terminal sequences. No reverse transcriptase product from this subclass has yet been demonstrated directly, although reverse transcriptase activity has been presumed from sequence homology with the other subclasses. In fact, homology among all three subclasses has been noted in the literature. However, no motifs for the class or subclasses are listed in the PROSITE database, and we were unable to find any motifs for any subclasses in the literature. The motif identification program in the original paper discovered the pattern `Y.DD`, and the literature mentions that `YMDD` and `Y[VLIM]DD` are common patterns.

We checked the identity and accuracy of the training set by comparing each sequence with the BLAST sequence alignment program [Altschul et al. 1990] against the SWISSPROT (version 30) and Protein Identification Resource (PIR, version 42) databases. One sequence, labeled ERVH in the original paper, could not be found in either database but was identified as HUMER41 by translation of nucleic acid sequences in the GENBANK database. Four sequences could not be found as written, but had close matches. Sequence POBO had an isoleucine in position 20, while its closest match, POL_BLVJ, had a tyrosine in that position. Likewise, sequences IFAC, INGT, and TYEY1 each had close matches with database entries B26330, S28721, and YCB9_YEAST, respectively, with the exception of three to four residues each. These mismatches may represent strain variants or, more likely, sequence errors. In particular, the four residue errors in IFAC could be explained by a deletion of a serine residue in position 4. Two sequences, HLIN and MLIN, were found to be poorly aligned. Their corresponding sequences, LIN1_HUMAN and POL2_MOUSE, each have a pair of adjacent aspartate residues which are aligned most appropriately at positions 16 and 17. Thus, we believe HLIN was aligned incorrectly at 7 residues too far to the left and mlin at 34 residues too far to the right. The correct alignments are shown in Table 1. The sequences in our example come from different subclasses of reverse transcriptases. Thirteen of the sequences are reverse transcriptases from retroviruses, eleven are class I retrotransposons, and six are class II retrotransposons. Of the remaining three sequences, two sequences are misaligned so much as to be uninformative, and one sequence, HEPB, is misclassified, since it is a DNA-directed DNA polymerase rather than an RNA-directed one.

Despite these errors, the sequences were given as input to SEQCLASS exactly as presented in the original paper. This enabled us to test the robustness of our algorithm to subclasses, contamination, and misalignment and sequence errors. The algorithm used a PAM value of 150 to measure the homology between sequences for ordering them. The resulting ordering is shown in Table 1. We set the beam width to 1000, and used the amino acid group database

Name	Sequence	Locus Name	Pos	Note	Source
TLPH	PMRKMFPTSTIVQYMDDILLASPTNEELQQ	POL_HTLV2	259	1	Pol Polyprotein [Human T-Cell Leukemia Virus Type II]
APOL	PIRQAFPQCTILQYMDDILLASPSHEDLLL	POL_HTL1A	174	1	Pol Polyprotein [Human T-Cell Leukemia Virus Type I]
HERV	PVREKFSDCYIIHYIDDILCAAETKDKLID	POL1_HUMAN	182	2	Retrovirus-Related Pol Polyprotein [Human]
RMTV	TVRDKYQDSYIVHYMDDILLAHPSRSIVDE	POL_MMTVB	174	1	Pol Polyprotein [Mouse Mammary Tumor Virus]
RVPO	PLRLKHPSLCMLHYMDDLLLAASSHDGLEA	POL_RSVP	167	1	Pol Polyprotein [Rous Sarcoma Virus]
HART	PIRKQFTSLIVIHYMDDILICHKELDVLQK	POL_IPHA	166	2	Putative Pol Polyprotein [Hamster Intracisternal A-Particle]
SADP	KVRHAWKQMYIIHYMDDILIAGKDGQQVLQ	POL_MPMV	180	1	Pol Polyprotein [Simian Mason-Pfizer Virus]
HIV2	PFRKANKDVIIIQYMDDILIASDRTDLEHD	POL_HV2RO	354	1	Pol Polyprotein [Human Immunodeficiency Virus Type II]
POBO	QVSAAFSQSLLVSYMDDILIASPTEEQRSQ Y	POL_BLVJ	148	1	Pol Polyprotein [Bovine Leukemia Virus]
HEPB	VVRRAFPHCLAFSYMDDVVLGAKSVQHLES	DPOL_HPBVP	538	?	DNA Polymerase [Hepatitis B Virus]
EAVP	PFRERYPEVQLYQYMDDLFVGSNGSKKQHK	POL_EIAVC	355	1	Pol Polyprotein [Equine Infectious Anemia Virus]
LVP1	DFRIQHPDLILLQYVDDLLLAATSELDCQQ	POL_MLVFF	334	1	Pol Polyprotein [Friend Murine Leukemia Virus]
RTAV	PFKKQNPDIVIYQYMDDLYVGSDLEIGQHR	POL_HV1B1	337	1	Pol Polyprotein [Human Immunodeficiency Virus Type I]
ERVH	KFPTRDLGCVLLQYVDDLLLGHPTAVGWPR	HUMER41	3634	2	Human endogenous retroviral DNA
COPD	DKGNINENIYVLLYVDDVVIATGDMTRMNN	COPI_DROME	1074	2	Copia Protein [D. melanogaster]
DIRS	LRMLRDINVSVIAYLDDLLIVGSTKEECLS	C24785	210	2	Transposon DIRS-1 Hypothetical Protein [Dictyostelium discoideum]
176D	DILRPLLNKHCLVYLDDIIVFSTSLDEHLQ	POL3_DROME	351	2	Retrovirus-Related Pol Polyprotein [D. melanogaster]
297D	NILRPLLNKHCLVYLDDIIIFSTSLTEHLN	POL2_DROME	350	2	Retrovirus-Related Pol Polyprotein [D. melanogaster]
GYPD	DVLREQIGKICYVYVDDVIIFSENESDHVR	POLY_DROME	326	2	Retrovirus-Related Pol Polyprotein [D. melanogaster]
412D	IAFSGIEPSQAFLYMDDLIVIGCSEKHMLK	POL4_DROME	459	2	Retrovirus-Related Pol Polyprotein [D. melanogaster]
HLIN	EVKLSLFADDMIVYLENPIVSAQNLLKLIS GIQLGKEEVKLSLFADDMIVYLENPIVSAQ	LIN1_HUMAN	693 686	>	Line-1 Reverse Transcriptase Homolog [Human] [Correct Alignment]
VLVP	GWIEEHPMIQFGIYMDDIYIGSDLGLEEHR	POL_VILV	311	1	Pol Polyprotein [Visna Lentivirus]
TYEY1	SCVFKNSQVTICLFVDDMVLFSKNLNSNKR I D A K	YCB9_YEAST	1004	2	Transposon TY1-17 154.0 KD Hypothetical Protein [S. cerevisiae]
CERV	NSHSNQYSKYCCVYVDDILVFSNTGRKEHY	POL_CERV	371	1	Enzymatic Polyprotein [Carnation Etched Ring Virus]
CAMV	DEAFRVFRKFCCVYVDDILVFSNNEEDHLL	POL_CAMVC	388	1	Enzymatic Polyprotein [Cauliflower Mosaic Virus]
TYS2	MADTFRDLRFVNVYLDDILIFSESPEEHWK	S41556	482	3	Retrotransposon TY3 Hypothetical Protein [S. cerevisiae]
QXB1	YNDPNFKRMKYVRYADDILIGVLGSKNDCK	YMX1_YEAST	509	3	Hypothetical COX1/OX13 Intron 1 Protein [S. cerevisiae]
IFAC	SNIILHKEIKFNAYADDFFLIINFNKNTNT NIIS	B26330	496	3	Transposon I factor [D. melanogaster]
C2IS	IHGIQSNKLMYVRYADDWIVAVNGSYTQTK	YM91_SCHPO	498	3	Hypothetical 91 KD Protein in COB Intron
INGT	QRLAEVPLLQHGFFADDLTLLARHTERDVI FS N	S28721	702	3	Hypothetical Protein 1 [Trypanosoma brucei]
MLIN	LKSGTRQGCPLSPYLFNIVLEVLARAIRQQ GIQIGKEEVKISLLADDMIVYISDPKNSTR	POL2_MOUSE	679 713	>	Retrovirus-Related Pol Polyprotein [Mouse] [Correct Alignment]
FE2D	SRSPIQATAQLALYLIDIKKWLSDWRIKVN	A32713	678	3	Reverse Transcriptase Homolog [D. melanogaster]
MAUP	VKELFKRYDELIMYADDGILCRQDPSTPDF	A25657	308	3	Hypothetical Mitochondrial Protein [Neurospora intermedia]

Table 1 Training set for reverse transcriptases. This example is taken from Smith HO and associates [1990]. The sequences are ordered according to the algorithm described in the text. Sequence errors are underlined and corrections shown below. The Pos column gives the starting position of the sequence. In the Note column, 1 signifies a retroviral reverse transcriptase; 2, a class I retrotransposon; 3, a class II retrotransposon; "?", a contaminating sequence; and ">", an alignment error.

Group	Elements	Source	Group	Elements	Source
a	AG	1	f	KR	1
b	ST	1	+	KRH	2
c	PAGST	2	g	VLI	3
d	QN	1	h	VLIM	1,2
–	ED	1	j	FYW	1,2
e	QNED	2	i	PAGSTQNEDHKR	1
			o	CVLIMFYW	1

Table 2 Database of amino acid groups. These amino acid groups were taken from evolutionary and functional studies by (1) [Jimenez-Montano & Zamora-Cortina 1981], (2) [Miyata, Miyazawa, & Yasunaga 1979], and (3) [Smith RF & Smith TF 1990]. In addition, the database contains singleton amino acid groups for each of the twenty amino acids.

shown in Table 2. We compiled the program using Lucid Common Lisp version 4.1.1 on a SUN SparcServer 1000 running the Solaris 4.3 operating system. The example required 253 seconds of CPU time.

Results

The ten highest-scoring motifs found by SEQCLASS are shown in Table 3. Each motif was constructed from a partition of the 33 sequences in the training set, and this partition is also shown for each motif. All of these high-scoring motifs excluded the sequences that were misaligned, namely, HLIN and MLIN. They also excluded the class II retrotransposons QXB1, IFAC, C2IS, INGT, FE2D, and MAUP. These retrotransposons all contain the pattern YADD at positions 14 through 17; apparently the program determined that generalizing the [VLIM] group at position 15 to include alanine would have reduced the significance of the motif too much. The class I retrotransposon TYEY1 was also generally excluded by our motifs, although motifs 8 and 9 did include it by generalizing the amino acid group at position 14 from [Y] to [FYW]. Other sequences that were excluded from at least one of the ten highest-scoring motifs were the class I retrotransposons HERV, 412D, and TYS2; the contaminating sequence HEPB; and the retroviral reverse transcriptase VILV. The ten motifs included all other retroviral reverse transcriptases. Note that the excluded sequences were generally ordered by the algorithm to be processed near the end. Thus, these sequences have a PAM-based homology that is generally distant from the other sequences. However, the homology criterion predicted the excluded sequences only approximately. For instance, although the misaligned sequence HLIN was excluded from all of the ten highest-scoring motifs, the retroviral sequence CAMV, which was judged to have less homology, was included in all of those motifs. In addition, the third sequence processed, the human sequence HERV, was excluded in three of the ten highest-scoring motifs.

We noted that our partitions generally included retroviral reverse transcriptases and class I retrotransposons, but excluded class II retrotransposons. Thus, we hypothesized that our motifs might identify these two subclasses of reverse transcriptases. To test this hypothesis, we assembled a "gold standard" list of retroviral and retrovirus-related reverse transcriptases. We used the IntelliGenetics program FINDSEQ to obtain all 137 sequences in SWISSPROT 30 that contained the annotation "RNA-directed DNA polymerase". We removed 29 sequences for which only fragments were known and 11 sequences that were class II retrotransposons, namely,

LIN1_HUMAN RT16_MYXXA RTJK_DROFU
LIN1_NYCCO RT65_MYXXA RTJZ_DROME
RDPO_SCEOB RT67_ECOLI RTP2_TRYBG
RRPO_OENBE RT86_ECOLI

One sequence, POLB_MAIZE, lacked any region of homology to the other sequences and was also deleted from the list. Another known class I retrotransposon, copia protein in Drosophila melanogaster (COPI_DROME), was also added to the list. The remaining 95 sequences are shown in Table 4. We noted that some sequences were duplicates over the region of 30 residues contained in our list. These sequences generally represented variations between different strains of the same organism. In order to avoid overweighting these duplicate sequences, we formed a second list of "distinct" retroviral and retrovirus-associated reverse transcriptases. In order to be considered distinct, a sequence had to differ from all other sequences in the list by at least one of the 30 residues in the region of interest. Thirty sequences were found to be duplicates, as shown in Table 4, leaving 65 distinct sequences.

For each of the ten highest-scoring motifs, we used the IntelliGenetics program QUEST [Abarbanel et al. 1984] to search for matching sequences in the SWISSPROT 30 database. For both the unmodified gold standard list and the distinct sequences, we tabulated true positives *TP* (sequences that both matched the motif and appeared in the gold standard list), false positives *FP* (those that matched the motif but did not appear in the gold standard list), and false negatives *FN* (those that appeared in the gold standard list but did not match the motif). Also, for the measures on the distinct sequences, we counted only distinct false positive hits. We used these quantities to compute the sensitivity (*TP*/(*TP*+*FN*)) and positive predictive value (*TP*/(*TP*+*FP*)) for each motif. False positive sequences were generally found to be viral DNA polymerases, most commonly DPOL_HPBVI, DPOL_WHV1, and their variants.

Table 5 shows that the motifs generated by SEQCLASS had reasonably high sensitivity and specificity. Motif 1 had a sensitivity of 91% and a positive predictive value of

	Signif	Motif	#Seqs	Sequences Excluded
		11111111112222	Incl	
		8901234567890123		
1	149.34	o..YhDDgoo.ii	22	HEPB, 412D, HLIN, TYEY1, QXB1, IFAC, C2IS, INGT, MLIN, FE2D, MAUP
2	149.34	i...o.YhDDgoo.i	22	HLIN, VLVP, TYEY1, TYS2, QXB1, IFAC, C2IS, INGT, MLIN, FE2D, MAUP
3	148.61	YhDDgoo.i	24	HLIN, TYEY1, QXB1, IFAC, C2IS, INGT, MLIN, FE2D, MAUP
4	147.46	i..oo.YhDDgoo.ii	20	HEPB, 412D, HLIN, VLVP, TYEY1, TYS2, QXB1, IFAC, C2IS, INGT, MLIN, FE2D, MAUP
5	146.58	YhDDgog.i	23	HERV, HLIN, TYEY1, QXB1, IFAC, C2IS, INGT, MLIN, FE2D, MAUP
6	146.26	o..YhDDgog.ii	21	HERV, HEPB, 412D, HLIN, TYEY1, QXB1, IFAC, C2IS, INGT, MLIN, FE2D, MAUP
7	146.26	i...o.YhDDgog.i	21	HERV, HLIN, VLVP, TYEY1, TYS2, QXB1, IFAC, C2IS, INGT, MLIN, FE2D, MAUP
8	146.14	i...o.jhDDhoo.i	23	HLIN, VLVP, TYS2, QXB1, IFAC, C2IS, INGT, MLIN, FE2D, MAUP
9	146.14	o..jhDDhoo.ii	23	HEPB, 412D, HLIN, QXB1, IFAC, C2IS, INGT, MLIN, FE2D, MAUP
10	146.13	YhDDgoo.ii	23	412D, HLIN, TYEY1, QXB1, IFAC, C2IS, INGT, MLIN, FE2D, MAUP

Table 3 Highest scoring motifs for reverse transcriptases. Lowercase motif symbols are explained in Table 2.

92% for distinct sequences. These values were slightly different for the unmodified gold standard list, due to weighting by multiple strains of the same organism. Motif 7 retrieved only half of the gold standard sequences. This gave it a low sensitivity but a high positive predictive value. This motif may perhaps identify a even more specific subclass of sequences. Initial analysis shows that it selects all retroviral sequences from HLTV-II, but none from HLTV-I.

For comparison, we tested motifs from the literature (results also shown in Table 5). The paper by Smith HO and associates [1990] that originally presented the example found the motif Y.DD. However, this motif is not specific. It retrieves over 1700 sequences in SWISSPROT 30, of which only a small fraction are reverse transcriptases, let alone subclasses of reverse transcriptases. Other researchers have noticed that position 15 is generally a small hydrophobic amino acid, but the motif Y[VLIM]DD still retrieves over 600 sequences. The motif YMDD, which occurs in most reverse transcriptases, increases the positive predictive value to 54% but lowers the sensitivity to 69%.

The insight provided by our motifs is that the region of consensus extends beyond the previously recognized four residues at positions 14 through 17. Our motifs all have a consensus at position 18, usually [VLI]. Furthermore, our motifs identify consensus positions of hydrophilicity and hydrophobicity. Within the motifs, position 11 or 12 is generally hydrophobic and positions 19 and 20 are always hydrophobic. Position 22 and sometimes position 23 are hydrophilic. These residues contribute greatly to the specificity of the motifs. To demonstrate their importance, we formed partial motifs, where the consensus positions for hydrophilicity and hydrophobicity were removed. As shown in Table 5, these motifs had much lower positive predictive values, indicating how important hydrophilic and hydrophobic residues are in identifying retrovirus and retroviral-related reverse transcriptases.

Discussion

The experiment carried out here demonstrates the power of conserved properties and partitioning methods in motif identification. Our program SEQCLASS discovered a novel motif for retroviral and retrovirus-related reverse transcriptases, a class of proteins for which no motif yet exists in the PROSITE database. Reverse transcriptases have received much scrutiny in biology and medicine because they play a critical role in the pathogenesis of acquired immunodeficiency syndrome (AIDS). Our motif exhibits regions of hydrophilicity and hydrophobicity outside the previously recognized region for reverse transcriptase motifs. These adjacent regions give the motif improved specificity and may be important in the structure of reverse transcriptases. These regions were found because our method allows multiple overlapping amino acid groups and can disregard incoherent data.

We anticipate several possible applications for our method. Given a training set, our method can identify a spectrum of motifs with varying levels of sensitivity and specificity. Alternatively, if the training set were assembled using a preliminary motif, our method could be used to refine or extend the motif. Our method may also be used to detect inconsistent data. Moreover, such apparently inconsistent data may indicate the presence of protein subclasses. Motifs for these subclasses might be identified by running our program subsequently on the set of excluded sequences.

In theoretical terms, our work adds new concepts to the field of motif identification. One concept is the need for robustness to errors, which are not uncommon in molecular biology. As a case in point, the example taken from the literature and used in this paper contains various errors. Probabilistic motifs and profiles may tolerate errors, but they allow these errors to degrade the result. Our method is not only tolerates errors, but also identifies and rejects them in order to obtain a more specific motif.

Another concept is the use of principled amino acid properties not merely to represent motifs but also to analyze the training set data for coherency. We effectively use knowledge about evolutionary, biological, and physical properties to help determine whether positions are conserved and whether sequences are incoherent. Our program currently only has a preliminary set of amino acid groups. The performance of our program may improve when other amino acid groups are added to the database [Kidera et al. 1985, Taylor 1986].

Locus Name	Pos	Sequence	Duplicate Sequences
COPI_DROME	1074	DKGNINENIYVLLYVDDVVIATGDMTRMNN	
POL1_HUMAN	182	PVREKFSDCYIIHYIDDILCAAETKDKLID	
POL2_DROME	350	NILRPLLNKHCLVYLDDIIIFSTSLTEHLN	
POL2_MOUSE	713	GIQIGKEEVKISLLADDMIVYISDPKNSTR	
POL3_DROME	351	DILRPLLNKHCLVYLDDIIVFSTSLDEHLQ	
POL4_DROME	459	IAFSGIEPSQAFLYMDDLIVIGCSEKHMLK	
POLR_DROME	566	GAKVGNAITNAAAFADDLVLFAETRMGLQV	
POLX_TOBAC	995	KRFSENNFIILLLYVDDMLIVGKDKGLIAK	
POLY_DROME	326	DVLREQIGKICYVYVDDVIIFSENESDHVR	
POL_BAEVM	327	DFRTQHPEVTLLQYVDDLLLAAPTKKACTQ	
POL_BIV06	326	NIKKSHPDVMLYQYMDDLLIGSNRDDHKQI	POL_BIV27
POL_BLVAU	148	QVSAAFSQSLLVSYMDDILYVSPTEEQRLQ	
POL_BLVJ	148	QVSAAFSQSLLVSYMDDILYASPTEEQRSQ	
POL_CAEVC	316	DWIQQHPEIQFGIYMDDIYIGSDLEIKKHR	
POL_CAMVC	388	DEAFRVFRKFCCVYVDDILVFSNNEEDHLL	POL_CAMVD, POL_CAMVE, POL_CAMVN, POL_CAMVS
POL_CERV	371	NSHSNQYSKYCCVYVDDILVFSNTGRKEHY	
POL_COYMV	1551	DNVFKGTEKFIAVYIDDILVFSETAEQHSQ	
POL_EIAV9	355	PFRERYPEVQLYQYMDDLFVGSNGSKKQHK	POL_EIAVC, POL_EIAVY
POL_FENV1	184	DFRTQHPEVTLLQYVDDLLLAAPTKEACIR	
POL_FIVPE	325	PFIRQNPQLDIYQYMDDIYIGSNLSKKEHK	POL_FIVSD
POL_FIVT2	324	PFIKQNSELDIYQYMDDIYIGSNLNKKEHK	
POL_FMVD	381	QTALNGADKFCMVYVDDIIVFSNSELDHYN	
POL_FOAMV	102	VVDLLKEIPNVQVYVDDIYLSHDDPKEHVQ	
POL_GALV	326	PFRALNPQVVLLQYVDDLLVAAPTYEDCKK	
POL_HTL1A	174	PIRQAFPQCTILQYMDDILLASPSHEDLLL	
POL_HTL1C	174	PIRQAFPQCTILQYMDDILLASPSHADLQL	
POL_HTLV2	259	PMRKMFPTSTIVQYMDDILLASPTNEELQQ	
POL_HV1A2	325	PFRKQNPDIVIYQYMDDLYVGSDLEIGQHR	
POL_HV1B1	337	PFKKQNPDIVIYQYMDDLYVGSDLEIGQHR	POL_HV1B5, POL_HV1BR, POL_HV1H2, POL_HV1MN POL_HV1N5, POL_HV1OY, POL_HV1PV
POL_HV1EL	324	PFRKQNPEMVIYQYMDDLYVGSDLEIGQHR	
POL_HV1JR	329	PFRKQNPDIIIYQYMDDLYVGSDLEIGQHR	
POL_HV1MA	324	PFRTKNPEIVIYQYMDDLYVGSDLEIGQHR	POL_HV1ND, POL_HV1RH, POL_HV1Z2
POL_HV1U4	324	PFRSQHPDIVIYQYMDDLYVGSDLEIGQHR	
POL_HV1Y2	325	PFRKQNPDLVIYQYMDDLYVGSDLEIGQHR	
POL_HV2BE	373	PFRKANPDVILIQYMDDILIASDRTGLEHD	POL_HV2D1, POL_HV2G1
POL_HV2CA	353	PFRKANSDVIIIQYMDDILIASDRTDLEHD	
POL_HV2D2	373	PFRKANSDVIIIQYMDDILIASDRSDLEHD	
POL_HV2NZ	353	PFRKANEDVIIIQYMDDILIASDRTDLEHD	
POL_HV2RO	354	PFRKANKDVIIIQYMDDILIASDRTDLEHD	
POL_HV2SB	353	PFRKANPDVIIVQYMDDILIASDRTDLEHD	
POL_HV2ST	373	PFRKANPDIILIQYMDDILIASDRTDLEHD	
POL_IPHA	166	PIRKQFTSLIVIHYMDDILICHKELDVLQK	
POL_IPMA	177	PVREQFPSLILLLYMDDILLCHKELTMLQK	
POL_IPMAI	96	PVREQFPSLILLLYMDDILLCHKDLTMLQK	
POL_JSRV	180	PVRQRFPQLYLVHYMDDILLAHTDEHLLYQ	
POL_MLVFF	334	DFRIQHPDLILLQYVDDLLLAATSELDCQQ	POL_MLVAV, POL_MLVF5, POL_MLVM, POL_MLVRD
POL_MMTVB	174	TVRDKYQDSYIVHYMDDILLAHPSRSIVDE	
POL_MPMV	180	KVRHAWKQMYIIHYMDDILIAGKDGQQVLQ	
POL_OMVVS	292	DWIAKHPMIQFGIYMDDIYIGSDLDIMKHR	
POL_RSVP	167	PLRLKHPSLCMLHYMDDLLLAASSHDGLEA	
POL_RTBV	1326	MQESFGDLKFALLYIDDILIASNNEKEHIE	POL_RTBVP
POL_SFV1	311	VVDLLKEIPNVQAYVDDIYISHDDPQEHLE	
POL_SFV3L	313	VVDLLKEVPNVQVYVDDIYISHDDPREHLE	
POL_SIVA1	361	EIKKELKQLTIVQYMDDLWVGSQEEGPKHD	
POL_SIVAG	366	EIKKELKPLTIVQYMDDLWVGSQEDEYTHD	
POL_SIVAI	364	EIKRHTPGLEIVQYMDDLWLASDHDETRHN	
POL_SIVAT	382	EIKRNLPALTIVQYMDDLWVGSQENEHTHD	
POL_SIVCZ	349	PFREKNPDITIYQYMDDLYVGSDLEIDQHR	
POL_SIVGB	334	VFRKNHPTVQLYQYMDDLFVGSDYTAEEHE	
POL_SIVMK	374	PFRKANPDVTLVQYMDDILIASDRTDLEHD	POL_SIVM1
POL_SIVSP	340	PFRKANPDVTLIQYMDDILIASDRTDLEHD	POL_SIVS4
POL_SMRVH	177	PVRSQWPEAYILHYMDDILLACDSAEAAKA	
POL_SOCMV	345	DQSLKGLDHIYLAYIDDILIFTKGSKEQHV	
POL_SRV1	180	KVRHAWKQMYIIHYMDDILIAGKDGQQVLQ	
POL_VILV	311	GWIEEHPMIQFGIYMDDIYIGSDLGLEEHR	POL_VILV1, POL_VILV2, POL_VILVK

Table 4 Gold standard list of retroviral and retrovirus-related reverse transcriptases. The SWISSPROT 30 locus names for distinct sequences are listed in the left column. Locus names for duplicate sequences over the region of 30 residues are listed in the right column.

Motif	All Sequences			Distinct Sequences		
	N	Sens	PPV	N	Sens	PPV
From the literature:						
Y.DD	1716	.98	.05	**	**	**
Y[VLIM]DD	609	.98	.15	**	**	**
YMDD	119	.69	.54	81	.69	.54
From SEQCLASS:						
o..Y[VLIM]DD[VLI]oo.ii	99	.93	.90	64	.91	.92
i...o.Y[VLIM]DD[VLI]oo.i	102	.83	.78	61	.82	.87
Y[VLIM]DD[VLI]oo.i	118	.96	.72	73	.94	.84
i..oo.Y[VLIM]DD[VLI]oo.ii	85	.82	.93	53	.80	.98
Y[VLIM]DD[VLI]o[VLI].i	115	.93	.77	70	.89	.83
o..Y[VLIM]DD[VLI]o[VLI].ii	96	.89	.92	58	.86	.97
i...o.Y[VLIM]DD[VLI]o[VLI].i	49	.46	.98	35	.52	.97
i...o.[FYW][VLIM]DD[VLIM]oo.i	104	.83	.77	62	.81	.85
o..[FYW][VLIM]DD[VLIM]oo.ii	101	.94	.89	66	.92	.91
Y[VLIM]DD[VLI]oo.ii	116	.95	.78	71	.92	.85
Partial (without hydrophilicity or hydrophobicity):						
Y[VLIM]DD[VLI]	203	.97	.45	158	.95	.39
Y[VLIM]DD[VLI].[VLI]	136	.94	.65	95	.91	.61
[FYW][VLIM]DD[VLIM]	337	.98	.27	291	.97	.21

Table 5 Sensitivity and positive predictive value of motifs. Legend: N = Number of sequences in SWISSPROT 30 that match the motif; Sens = Sensitivity; PPV = Positive predictive value; ** = Too many matches to identify distinct sequences.

The techniques of using principled amino acid groups and identifying incoherent data work synergistically. Coherent data indicates the constraints that may hold in a protein, and conversely, those constraints indicate which data may be incoherent. Moreover, our work illustrates the interplay between finding patterns *within* sequences and finding patterns *of* sequences. Protein sequence data that are incoherent often appear random and unstructured. It is only when coherent patterns of sequences are assembled and analyzed that specific patterns within the sequences appear.

References

Abarbanel, R. M., Wieneke, P. R., Mansfield, E., Jaffe, D. A., and Brutlag, D. L. 1984. Rapid searches for complex patterns in biological molecules. *Nucleic Acids Research* 12:263–280.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215(3):403–410.

Bairoch, A. 1991. PROSITE: A dictionary of sites and patterns in proteins (release 6.1). *Nucleic Acids Research* 19(Supplement):2241–2245.

Gribskov, M., Luthy, R., and Eisenberg, D. 1990. Profile analysis. *Methods in Enzymology* 183:146–159.

Henikoff, S. and Henikoff, J. G. 1991. Automated assembly of protein blocks for database searching. *Nucleic Acids Research* 19(23):6565–6572.

Jimenez-Montano, M. A., and Zamora-Cortina, L. 1981. Evolutionary model for the generation of amino acid sequences and its application to the study of mammal alpha-hemoglobin chains. In Proceedings of the Seventh International Biophysics Congress, Mexico City.

Jones, D. T., Taylor, W. R., and Thornton, J. M. 1992. The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences* 8(3):275–282.

Kidera, A., Yonishi, Y., Masahito, O., Ooi, T., and Scheraga, H. A. 1985. Statistical analysis of the physical properties of the twenty naturally occurring amino acids. *Journal of Protein Chemistry* 4:23–55.

Miyata, T., Miyazawa, S., and Yasunaga, T. 1979. Two types of amino acid substitution in protein evolution. *Journal of Molecular Evolution* 12:219–236.

Saqi, M. A. S. and Sternberg, M. J. E. 1994. Identification of sequence motifs from a set of proteins with related function. *Protein Engineering* 7(2):165–171.

Smith, H. O., Annau, T. M., and Chandrasegaran, S. 1990. Finding sequence motifs in groups of functionally related proteins. *Proceedings of the National Academy of Sciences* 87(2):826–830.

Smith, R. F. and Smith, T. F. 1990. Automatic generation of primary sequence patterns from sets of related protein sequences. *Proceedings of the National Academy of Sciences* 87:118–122.

Tatusov, R. L., Altschul, S. F., and Koonin, E. V. 1994. Detection of conserved segments in proteins: Iterative scanning of sequence databases with alignment blocks. *Proceedings of the National Academy of Sciences* 91:12091–12095.

Taylor, W. R. 1986. The classification of amino acid conservation. *Journal of Theoretical Biology* 119:205–218.

Wu, T. D. 1990. Efficient diagnosis of multiple disorders based on a symptom clustering approach. In Proceedings of the Eighth National Conference on Artificial Intelligence, 357–364. Menlo Park, Calif.: AAAI Press.