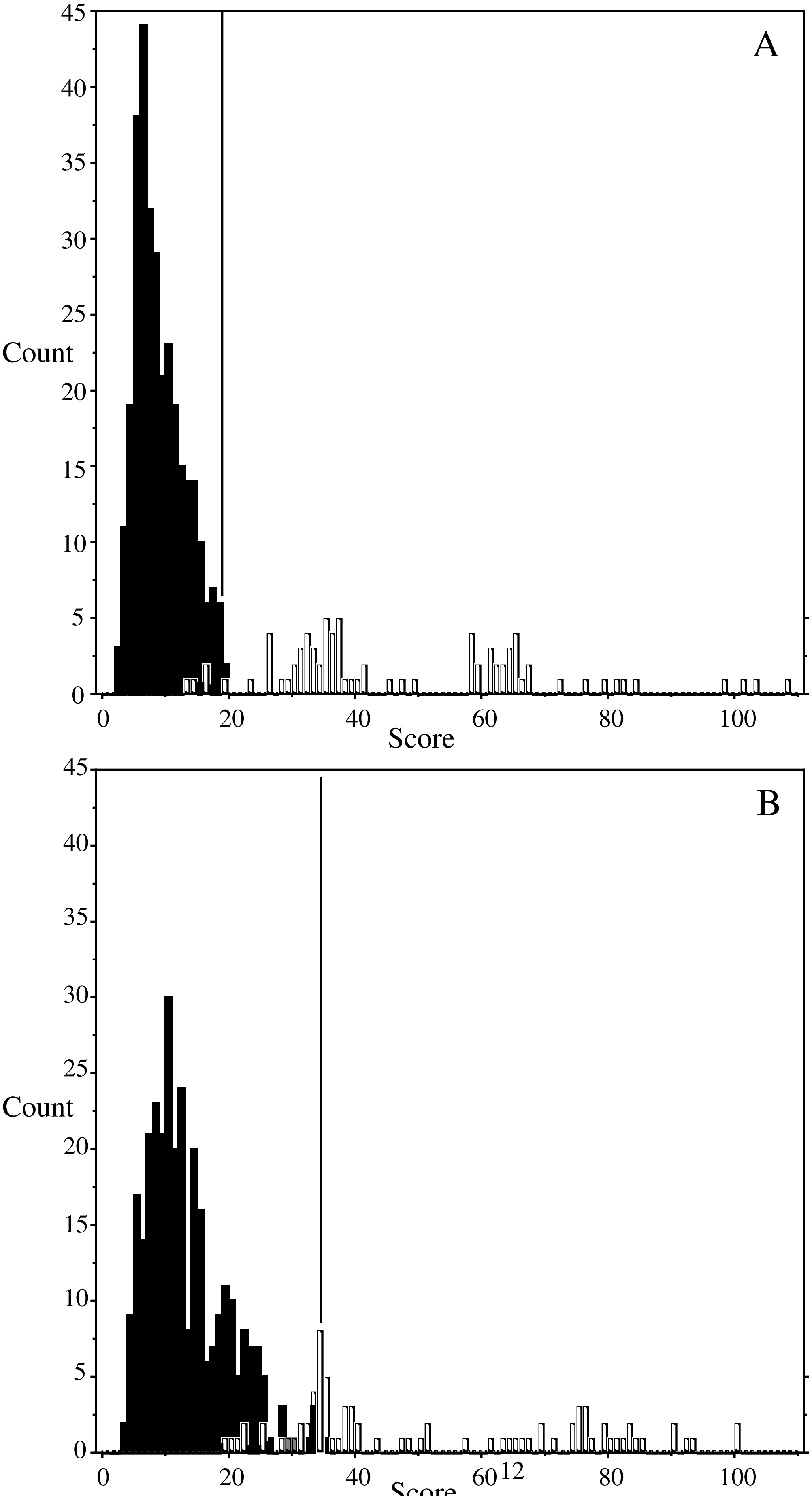
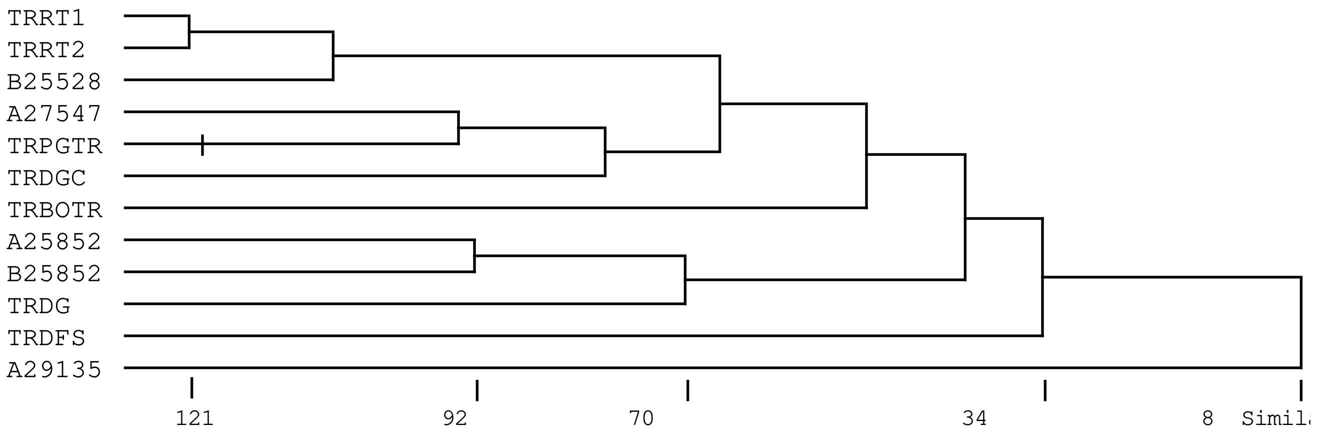
Fig. 5 (prev. page). Match score distributions comparing AACC pattern vs. sequence matches (A) to sequence vs. sequence matches (B). (A) An AACC pattern was constructed from 4 protein tyrosine kinase sequences (Fig. 4D) and matched to 2 control sets: a positive control set composed of 81 protein kinase (PK) related sequences (white bars) and a negative control set 313 non-PK-related sequences (black bars; mean ± s.d = 10.88 ± 4.60); where the two sets of matches overlap, the smaller bar at each position is shown in the foreground; at those positions where the bars are the same height, a single stippled bar is displayed. The dashed line indicates the 99th percentile of the negative control scores. (B) Score distributions of a single PK sequence (human *ret*, locus TVHURE) matched against the same positive (white bars) and negative (black bars; mean ± s.d = 13.59 ± 6.59) control sets described above. As in (A), the dashed line indicates the 99th percentile of the negative control set. The match score of TVHURE vs. TVHURE is 540.6; this sequence was not included in the positive control set since it is labelled as a "fragment" in the database.


A
Count
0
5
10
15
20
25
30
35
40
45
0
20
40
60
80
100
Score
B
Count
0
5
10
15
20
25
30
35
40
45
0
20
40
60
80
100
Score

A. Cluster 50 (Trypsinogen/Venom serine proteases)



TRRT1	Trypsinogen	I precursor	- Rat	#EC-number 3.4.21.4
TRRT2	Trypsinogen	II precursor	- Rat	#EC-number 3.4.21.4
B25528	Trypsinogen	precursor	- Mouse	#EC-number 3.4.21.4
A27547	Trypsinogen,	cationic, precursor	- Rat	#EC-number 3.4.21.4
TRPGTR	Trypsinogen		- Pig	#EC-number 3.4.21.4
TRDGC	Trypsinogen	cationic, precursor	- Dog	#EC-number 3.4.21.4
TRBOTR	Trypsinogen		- Bovine	#EC-number 3.4.21.4
A25852	Trypsinogen	I precursor	- Human	#EC-number 3.4.21.4
B25852	Trypsinogen	II precursor	- Rat	#EC-number 3.4.21.4
TRDG	Trypsinogen	anionic, precursor	- Dog	#EC-number 3.4.21.4
TRDFS	Trypsinogen		- Spiny Dogfish	#EC-number 3.4.21.4
A29135	Venom serine protease		- Barba amarilla	#EC-number 3.4.21.29

B. Pattern 50 (from cluster 50, above)

1 20 21 40 41 60 61 80
lllhXaaGGXX**C**XXlXXPbX XXcXXXXigF**C**GXkLIXXXW VakAp**HC**XlXXclaiLGlXX XXXXlXXEXcXXXXXXcXXP
XXXlXlaXXaXLPlXXXXXG lXXXaXGWGXXXlgXXXXXl XXl**C**XlXXacXlXX**C**lXXYX GgaXXlXc**C**XGcclGGX**DkC**
GXXg**C**AXXXXPpcXXiVclb aXWIllXaA

C. Pattern 1 (from cluster 1, Table 1: Hemoglobin beta, delta, epsilon, gamma)

llXXaXXXbXXcXXXXXGXl XaXlccaac**P**WllRb**F**XXFG XcXlXXXaXXlXXaXXXGXi aXXXcXXXcXlclXaXXXcX
GXXcaXXcXXXXXXXFXXXX lXXclicXXXaXXpLXXXY

D. Pattern constructed from 4 protein tyrosine kinase sequences

lXaXXaXXXaXXcXXXXlXX XXXXXlXgggXlXlXlllXl ghlAgXlcXXcXgggXXXXl laXXXXXXXXXXXXXXlXXX
XggggggXXXXXXlXXlXXX lXXXXXXXXXiXXaXXXGXX XXlXXiXlXXXXXXXgglXl XXaXXcXXXXXXXXXXXllc
XlXXlXXgggggggglXXlgg lXXXXXXXTXXXXXXlXXXX XlXXXXXXXXXXXXXXXaXX XXXXXXXXXXcaXXXXXcXX
XXXXXcXbXlXXXXgglXXX XXXXXSXXXXXaXXlXXXgg gggggggglcXRllaXLXXl **LGlGXFGXV**cXplXXXaXXX
XXcXlEXlaclXcXgXXlaa lccGXcklXXPXccaXEcXX XGlLXXbLilggggggggg gggggggggggggggggg
ggggggggggggggggggg ggggggggggggggggggg ggggggggggXXXXcXlcaX cXXlaXXGMXbLXXXl**cVHR**
GcXRDaclllXYXXXXlXX**a PaiWMXXE**SaXXXXbTTXS**D cWSFGaaLWEIXkcpXlPb**X XaXXllXcglXaXlGXXcXl
lXggggXXXXlXcXXXXggi XXXXXcXXXXXlXXXXXXlX XXgggggXXXXlXXXcP

	A	B	C.1	C.2
seq/pat 1	`ADEFGHIK`	`ADggggIK`	`ADEFGHIK`	`ADggGHIK`
seq/pat 2	`AD----IK`	`AD----IK`	`ADgg--IK`	`AD----IK`
	⇓	⇓	⇓	⇓
pattern	`ADggggIK`	`ADggggIK`	`ADggggIK`	`ADggggIK`
`gap penalty`	$W= W_1+W_2*4$	$W= 0$	$W= W_2*2$	$W= W_2*2$

Fig. 3. Four examples of the application of the "pay once" rule for gap weighting during sequence/pattern alignment. Sequence characters are as defined in Fig. 1; a dash ("-") represents a single residue gap created during alignment. In the 4 examples, two sequence/pattern fragments are shown aligned (top); the resulting pattern is shown below.

Fig. 4 (next. page). (A) A cluster of serine protease-related sequences generated by clustering the NBRF/PIR database (cluster 50). (B) The AACC pattern generated from cluster 50. Upper case characters: IUPAC one-letter amino acid codes; lower case characters: amino acid classes (see Fig. 1) a=[ILV], b=[FWY], c=[ILVFWYCM], h=[DE], i=[HKR], j=[NQ], k=[ST], l=[DEHKRNQSTBZ], p=[AG], X=wild-card (1 amino acid of any type), g= gap character (0 or 1 amino acid of any type). The 8 cysteine residues forming the 4 disulfide bonds that are common to this set of proteases (17-19) are all present in the pattern and are shown highlighted in bold/underlined type. The highlighted regions at positions 178-188 and 199-200 are at the substrate binding site (20); the former region contains the "GDSG" motif conserved in known serine proteases (21). Highlighted positions 46 and 91, along with position 184 constitute the active site charge relay system (20). (C) AACC pattern generated from cluster 1 (composed of hemoglobin beta, delta, epsilon and gamma sequences; see Table 1). Three required or highly conserved residues within globin sequences (22) are shown highlighted: the "proximal" histidine residue in the F helix which binds iron (pos. 86), a phenylalanine residue in the C helix that packs against the heme (pos. 36), and a proline residue in the C helix (pos. 30). The "distal" ligand binding histidine found in most globins at pos. 57 is not conserved in this sequence set. (D) AACC pattern constructed from 4 protein serine/threonine kinase sequences: human insulin receptor (NBRF/PIR locus A05274) and feline *trk* (locus A25184), mouse PDGF receptor (locus A25742) and human *ret* (locus TVHURE). Six conserved regions previously identified within the catalytic domains of protein kinases (23, 24) are shown highlighted.

FIGURES

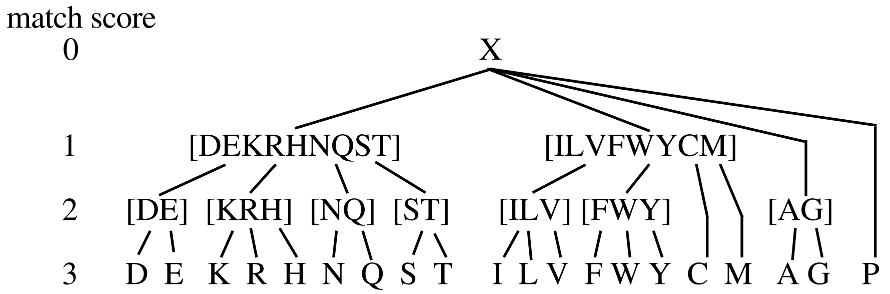


Fig. 1. The amino acid class hierarchy used to construct the amino acid class covering (AACC) patterns; Upper case characters: IUPAC one-letter amino acid codes; X: wild-card character (1 amino acid of any type); g: gap character (0 or 1 amino acid of any type).

Pat-1		Pat-2	Match score	Covering pattern	Single letter code
D	–	D	3	D	D
D	–	E	2	[DE]	h
D	–	[DE]	2	[DE]	h
[DE]	–	[DE]	2	[DE]	h
[DE]	–	[KRH]	1	[DEKRHNQST]	l
[DE]	–	X	0	X	X
[DE]	–	g	0	g	g
etc		.	.	.	.

Fig. 2. An example of an alignment between two previously constructed patterns (designated Pat-1 and Pat-2), showing 1) the resulting match scores between aligned pattern elements, 2) construction of a covering pattern from the minimally inclusive class covering both aligned elements (see Fig. 1), and 3) the final pattern as a sequence composed of upper and lower case characters representing individual amino acids (e.g. D, E), amino acid classes (e.g. h= [DE], l= [DEKRHNQST], X, etc.; see Fig. 4 legend), and gapped positions (g).

16. Wilbur, W.J. & Lipman, D.J. (1983) Proc. Natl. Acad. Sci. USA. **80**:726-730.
17. Mikes, O., Holeysovsky, V., Tomasek, V. & Sorm, F. (1966) Biochem. Biophys. Res. Commun. **24**:346-352.
18. Emi, M., Nakamura, Y., Ogawa, M., Yamamoto, T., Nishide, T., Mori, T. & Matsubara, K. (1986) Gene **41**:305-310.
19. Itoh, N., Tanaka, N., Mihashi, S. & Yamashina, I. (1987) J. Biol. Chem. **262**:3132-3135.
20. Bode, W. & Schwager, P. (1975) J. Mol. Biol. **98**:693-717.
21. Leytus, S.P., Loeb, K.R., Hagen, F.S., Kurachi, K. and Davie, E.W. (1988) Biochemistry. **27**:1067-1074.
22. Bashford, D., Chothia, C. & Lesk, A.M. (1987) J. Mol. Biol. **196**:199-216.
23. Hanks, S.K., Quinn, A.M., & Hunter, T. (1988) Science **241**:42-52.
24. Smith, R.F. & Smith, T.F. (1989) J. Virol. **63**, 450-455.
25. Smith, T.F., Waterman, M.S. & Burks, C. (1985) Nucl. Acids Res. **13**:645-657.
26. Reichardt, J.K.V. & Berg, P. (1988). Nucl. Acids Res. **16**:9017-9026.

distributions of the positive and negative control sets.

It is possible to generalize the patterns constructed from sequence families in the NBRF/PIR database by combining clusters from related sub-families (e.g. alpha + beta + myo-globins) within larger "super-families" (e.g., globins) but only if sufficient information content (IC) can be maintained in the final pattern. The pattern construction methodology is of course limited by the degree of primary sequence similarity shared by any set of input sequences. If a set of sequences used to construct a pattern has no conserved primary sequence elements in common, then the final "root" pattern will have little if any information content over that expected by chance. In general, patterns with low IC will have considerable overlap between positive and negative control distributions and therefore will have limited diagnostic ability. In the set of AACC patterns that we have created from the NBRF/PIR database, all have IC of 10 or greater, and all have clear separation between the score distributions of the pattern vs. 1) the negative control set and 2) the set of sequences from which the pattern was generated.

AACC patterns generated from sequence families contain conserved regions that include amino acids that are likely to be functionally important. For example, such previously defined sites as the nucleotide binding site of the protein kinases (GxGxxG...AxK; ref. 23) are clearly identifiable within the PK pattern of Fig. 4D. The identification of patterns representing conserved amino acid sites/domains should prove useful in gene cloning, protein engineering, and protein function studies (26). However, those elements found to be shared among any family of sequences represent two classes: those that are functionally conserved and those that by chance have not yet varied within whatever constraints may apply in regions not directly involved in the particular function. Thus patterns derived from a set of related sequences will identify functionally important sites/domains only if significant sequence divergence (e.g. broad taxonomic breadth) is represented within the clustered set.

We thank Karen Kabnick and Teresa Webster for critical reading of the text, David Lipman and Sam Karlin for helpful discussions, and Don Faulkner for programming assistance. This work was supported by Public Health Service grant RR02275 from the National Institutes of Health.

1. Hodgman, T.C. (1986) CABIOS **2**:181-187
2. Taylor, W.R. (1986) J. Mol. Biol. **188**:233-258.
3. Blundel, T.L., Sibanda, B.L., Sternberg, M.J.E. & Thornton, J.M. (1987) Nature **326**:347-352.
4. Gribskov, M., McLachlan, A.D., & Eisenberg, D. (1987) Proc. Natl. Acad. Sci. USA. **84**:4355-4358.
5. Patthy, L. (1987) J. Mol. Biol. **198**: 567-577.
6. Zvelebil, M.J., Barton, G.J., Taylor, W.R. & Sternberg, M.J.E. (1987) J. Mol. Biol. **195**:957-961.
7. Pearson, W.R. & Lipman, D.J. (1988) Proc. Natl. Acad. Sci. USA **85**:2444-2448.
8. Taylor, W.R. (1988) Protein Engineering **2**:77-86.
9. George, D.G., Barker,W.C., & Hunt, L.T. (1986) Nucl. Acids Res. **14**:11-15.
10. Smith, T.F. & Waterman, M.S. (1981) J. Mol. Biol. **147**:195-197.
11. Smith, T.F. & Waterman, M.S. (1981) Adv. Appl. Math. **2**:482-489.
12. Waterman, M.S. (1984) J. Theor. Biol. **108**:333-337.
13. Aho, A. (1980) in *Formal Language Theory*, ed. R. Book (Academic Press, New York).
14. Arbarbanel, R.M., Wieneke, P.R., Mansfield, E., Jaffe, D.A. & Brutlag, D.L. (1984) Nucl. Acids Res. **12**:263-280.
15. Sneath, P.H. & Sokal R.R. (1973) *Numerical Taxonomy* (W.H. Freeman, San Francisco).

only one member) and excluding known PK sequences.

Length-normalized scores were generated by optimally aligning the PK pattern to both the positive and negative control sets (Fig. 5A). False-negative (FN) matches were considered those matches within the positive control set of PK sequences with scores less than the 99th percentile of the match scores of the non-PK-related negative control set. For the PK pattern, only 4 PK sequences had match scores less than the 99th percentile of control set (Fig. 5A). The 4 PK sequences used to generate the pattern were also individually aligned with the same two control sets. Using the same 99% percentile criterion for determining FN matches, the 4 individual PK sequences (human insulin receptor, feline *trk*, mouse PDGF receptor, and human *ret*) generated 46, 23, 29, and 20 FN matches, respectively. The positive and negative score distributions for the PK sequence with the least number of FN matches, human *ret*, are shown in Fig. 5B. The greatly reduced level of FN matching for the PK pattern vs. the individual PK sequences demonstrates that AACC patterns can be more diagnostic of family membership than the sequences used to construct the pattern.

DISCUSSION

We have used the same amino acid classification scheme to define both the AACC pattern elements and a similarity scoring metric among these same elements. Given this particular similarity matrix and gap scoring scheme, for any single pair of amino acid sequences and/or patterns, the present method will find the optimal AACC pattern (i.e. one that has the the highest score of any such pattern when matched against the two generating sequences. However, since we employ sequential pairwise alignments, rather than a single optimal multi-sequence alignment, during pattern construction, our method cannot guarantee that the final (root) pattern is optimal for the entire set. Nevertheless, given that each of the pairwise alignments contains the maximal set of conserved elements common to the entire set, even if the pairwise alignments are not all equivalent (different gap arrangements extending to different degrees between common domains), the method should find the overall optimal pattern for the set. In fact, since the patterns we have tested are not very sensitive to cluster branching (and thus to alignment order), the sequential pairwise alignment technique appears to generate nearly optimal patterns in practice.

There are two apparent reasons as to why AACC patterns can be more diagnostic of family membership than the sequences used to construct them. First, in families such as the protein kinases, mismatch and gap penalties generated at non-conserved positions by sequence vs. sequence alignments can easily outweigh the match scores contributed by the few, dispersed conserved positions. Second, since non-conserved sites/regions within sequences are not represented in the patterns, chance similarities between non-conserved sites within non-homologous sequences are eliminated. Such chance similarities would broaden the score distribution of the negative control set increasing the potential overlap between the score generated using just identities but include additional pattern elements apparently created by chance coincidence with one of the random classes, resulting in higher IC patterns.

To determine the degree to which final patterns are sensitive to the branching order of the cluster, we carried out two empirical tests. First, patterns were generated from subsets of sequences taken from a cluster and then compared to the pattern generated from the entire family. In the subsets, the degree of pattern degeneracy was directly related to the degree of sequence divergence represented in the subset; if a subset sampled the complete taxonomic range of the original cluster, then the resulting pattern closely resembled the pattern generated from the complete set. Second, patterns were generated from clusters in which sequences were randomly reassigned to the tree termini and then compared to the pattern generated from the original cluster. In all cases the patterns generated from clusters with randomly assigned termini closely resembled the pattern generated from the original cluster hierarchy. Thus, given the degree of sequence relatedness that was used to define sequence families during clustering, the final patterns are relatively insensitive to branching order and therefore to the initial set of pairwise scores.

Evaluating the diagnostic capability of AACC patterns. By comparing the diagnostic capability of a pattern with that of the sequences used to construct the pattern, we found that AACC patterns can be more diagnostic than the individual sequences in detecting distantly related family members . In these studies, the same dynamic programming method, including gap weighting function, was used, but "mismatches" were more strongly penalized (-2 instead of 0 except for matches to ’X’ which were maintained at 0). The scores were length-normalized by dividing by the logarithm of the product of the lengths of the aligned sequences (effectively dividing by the similarity score expected by chance; ref. 25); here it is assumed that the similarity scores obtained using our gap penalty and weighting matrix have the same length dependence as noted earlier for DNA comparisons (25). For the AACC patterns we have used the information content, measured in amino acid equivalents, as their effective lengths (for sequences, sequence length is equal to the information content when expressed as amino acid equivalents).

The protein kinase (PK) family was used to evaluate the diagnostic capability of AACC patterns since this family encompasses a large and diverse set of sequences with low overall sequence similarity (23, 24). An AACC pattern (Fig. 4D) was constructed from 4 protein tyrosine kinase sequences (2 from the insulin receptor subfamily, human insulin receptor and feline *trk*, and 2 from the PDGF receptor subfamily, mouse PDGF receptor and human *ret*) (23). A positive control set of 80 PK sequences was then constructed by 1) searching for the words "protein kinase" in the comment fields of the sequences in the NBRF/PIR database, and 2) using a regular expression sequence pattern that detects most known PK sequences in the database (pattern III.B of ref. 24), and excluding all sequence fragments. One additional known PK-related sequence that is not present in the NBRF/PIR database (EBV gene BGLF4) (17) was also added to this set. A negative control set of 313 non-PK-related sequences was generated by taking one sequence from every tenth sequence family generated from the NBRF/PIR database (including those clusters with where $|AAC_i|$ is the cardinality of the i^{th} amino acid class in the pattern and n_j is the length of the j^{th} gap in the pattern. Note that IC is not defined if the expectation value exceeds unity, which can happen for patterns violating assumption two above. The assumption of informational equivalence of the 20 amino acids is not meant as a statement about amino acid occurrence frequencies. Rather this assumption allows each pattern to be assigned a relative measure independent of amino acid composition.

RESULTS

Constructing AACC patterns for all NBRF/PIR sequence families. Our pattern construction methodology was applied to all families of closely related sequences in the NBRF/PIR v.18 database (9). To enable the entire database to be clustered efficiently, pairwise match scores between all sequences in the database were generated using a high-speed k-tuple/hash algorithm (DASHER; D.V. Faulkner and T.F. Smith, Molecular Biology Computer Research Resource, 1987) similar to the Wilbur-Lipman algorithm (23). A k-tuple of 3 and a reduced amino acid alphabet of 13 was used in these comparisons. Since patterns can be no longer than the shortest parental sequence, all sequences labeled as a "fragment" in their NBRF/PIR title lines were excluded (1743 loci). In this study, sequence families were conservatively defined as those with DASHER Chi-squared similarity scores of 4.0 or greater (encompassing the largest 0.2% of all of these pairwise scores); this cutoff was chosen so that the alpha globins would just separate from beta globin sequences. Using this cutoff, the 7395 non-fragment sequences in the database clustered into 1215 families of 2 sequences or more, encompassing 5321 of the 7395 sequences.

Next, an AACC pattern was constructed for each of the 1215 sequence families. The 12 largest clusters and their corresponding AACC patterns are profiled in Table 1. Three sample patterns (from cluster 50, a serine protease-related family; from cluster 1, the largest cluster, composed of hemoglobin beta and related delta, epsilon and gamma sequences; and from a cluster of 4 protein tyrosine kinase sequences [see below]) are shown in Fig 4B-C. Functionally conserved residues and sub-domains that have been previously identified in these families (17-24) are highlighted and are described in the figure legend.

Evaluating the effects of the amino acid class hierarchy and cluster topology on pattern generation. The amino acid hierarchical classification scheme that we have used (Fig.1) , based on side-chain physiochemical properties, clearly does not allow for all known conservative substitutions. To evaluate the sensitivity of this classification, patterns were constructed for 36 sequence clusters using 1) no amino acid classes beyond identities, 'X', and 'g', and 2) 5 classifications each constructed by randomly assigning the 20 amino acids to the termini of the hierarchy in Fig. 1. Using only amino acid identities resulted in a 38% average decrease in the IC values compared to our standard scheme (a reduction of about 23 amino acid equivalents per pattern). The 5 random classification schemes on average resulted in a 12% decrease in the IC values compared to our standard hierarchy. The latter patterns resemble those as the penalty for creating or initiating a gap of any length, and should be imposed only once and not for the subsequent lengthening of a pre-existing gap. Provisions 2 and 4 allow gap characters to function as 0 or 1 amino acids of any type (analogous to the way variable spacing can be specified in regular expression patterns, refs. 13, 14), and ensure that the longest gap observed in any pattern will be preserved. A pattern will thus have an identical match score when aligned with any of the "parental" sequences from which the pattern was derived, again in a manner analogous to a regular expression.

Covering Pattern Construction. An amino acid class covering (AACC) pattern is generated from any given pairwise alignment by the following prescription. If any amino acid is aligned with itself, that amino acid is placed in the pattern at the equivalent position. When any amino acid or class of amino acids is aligned with a different amino acid or class, the minimally inclusive class covering both aligned elements (Fig. 1) is placed at that position (Fig. 2). When any character is aligned with a previously introduced gap, a gap character is placed at that position (Figs. 2, 3C.1). At any position within the alignment at which a new gap has been inserted in either sequence, a gap character ('g') is inserted in the covering pattern for each residue spanning the gap (Fig. 3). In any region of the alignment where there are alternate optimal paths, the extreme path placing all gap characters as a group at the N-terminal part of the degenerate region was chosen. In the patterns, individual amino acids are designated by upper case IUPAC single letter codes; amino acid classes are each designated by a single specified lower case character (Figs. 2, 4B-C).

To construct a single (or "root") overall covering pattern for a set of related protein sequences, all pairwise comparisons between sequences in the set are first performed. The pairwise scores are then clustered into a binary dendrogram (Fig. 4A), using a maximum linkage rule (15). Beginning with the most closely related pair of sequences in the cluster, the maximal local similarity alignment and associated covering pattern are obtained as described above. The two aligned sequences are then replaced in the cluster hierarchy by the AACC pattern at the similarity level of their common node. The alignment and pattern construction steps are then iteratively performed for each remaining node in decreasing order of similarity until a single covering pattern replaces the root node. The root covering pattern generated in this fashion will match with equal score all of the parental sequences from which the pattern was derived.

Calculation of the Relative Information Content of Patterns. Under two simplifying assumptions one can assign a relative information content (IC) measure to the AACC patterns: first, that all amino acids are informationally equivalent; and second, that those regions separated by gaps are statistically independent. The information content in amino acid equivalents is then defined as the logarithm to the base 20 of an expectation value:

$$\mathrm{IC} = -\log_{20}\left\{\left(\prod_{i}\left(|\mathrm{AAC}_i| * 1/20\right)\right)\prod_{j}\left(n_j + 1\right)\right\} \qquad (1)$$

One of the major problems in molecular biology is the identification of essential amino acid sequence elements that encode functional domains of proteins. While X-ray structure analysis of protein-substrate co-crystals is the most direct method, sequence comparative methods are far easier and are routinely used to identify conserved sites/regions within a set of homologous sequences which are likely to be functionally important (1-8). Ideally, one would like a simple method for generating the pattern of sequence elements essential for a given protein functional domain. Such patterns may be highly dispersed at the primary sequence level since protein folding can bring distant primary sequence elements into close three-dimensional proximity. If sufficiently complex, these patterns would be diagnostic of membership within a particular homologous family or functional class. As an initial step in this direction, we have developed a method to extract a diagnostic pattern of primary sequence elements from any set of protein sequences displaying a minimum overall group similarity. We have used this method to construct a full set of diagnostic sequence patterns for all protein families in the NBRF/PIR (9) protein sequence database.

METHODS

Dynamic Programming Alignments. We have extended the dynamic programming algorithm of Smith and Waterman (10) to generate sequence patterns from locally optimal pairwise alignments. Our modification enables the resulting pattern to be used as an input sequence for subsequent pairwise alignments. The sequence element-to-element similarity matrix used was generated on an alphabet composed of the 20 amino acids plus 10 amino acid classes, the latter based on the physiochemical nature of the side groups. These classes form a simple 4-level nested hierarchy (Fig. 1) that includes a "wild card" class, 'X', representing an amino acid of any type. This class hierarchy is also used to determine the minimal amino acid class covering any two aligned elements as performed during pattern construction (described below). The values of the similarity matrix elements range from +3 to 0 (see Fig. 1), with 0 implying maximum dissimilarity; the match score between any two aligned elements equals the score assigned to the minimally inclusive class in the hierarchy that covers both elements (see Fig. 2).

Our major extension to the standard dynamic programming algorithm involves the development a new "pay once" rule for gap weighting. Here, the gap penalty value has been made a function of whether there has been a gap previously inserted in a pattern to be aligned. Our "pay once" rule has four provisions: 1) for a gap of length k, a length-independent plus a length-dependent gap penalty ($W_1 + W_2*k$) (11,12) is imposed upon the initial introduction of a gap into an alignment (Fig. 3A); 2) no gap penalty is applied for the insertion of a single residue length gap across from a previous gapped position (represented as a gap character, 'g') in a pattern (Fig. 3B); 3) only the length-dependent penalty, W_2*k, is applied when extending a gap adjacent to a gap character in either of the sequences being aligned (Fig. 3C.1-2); and 4) the alignment of a gap character with any other character has a match score of 0. The motivation for provisions 1-3 is that W_1 is viewed

Classification: Biochemistry PNAS 87:118-122 (1990)

AUTOMATIC GENERATION OF PRIMARY SEQUENCE PATTERNS FROM SETS OF RELATED PROTEIN SEQUENCES

(amino acid/sequence comparison/protein families/ cluster analysis/dynamic programming)

Randall F. Smith[*] and Temple F. Smith

Molecular Biology Computer Research Resource, Department of Biostatistics, Dana-Farber Cancer
Institute and School of Public Health, Harvard University, Galleria Level 1,
44 Binney St., Boston, MA. 02115 USA

Communicated by: Samuel Karlin, September 18, 1989.

ABSTRACT We have developed a new computer algorithm that can extract the pattern of conserved primary sequence elements common to all members of a homologous protein family. The method involves clustering the pair-wise similarity scores among a set of related sequences to generate a binary dendrogram (tree). The tree is then reduced in a step-wise manner by progressively replacing the node connecting the two most similar termini by one common pattern until only a single common "root" pattern remains. A pattern is generated at a node by 1) performing a local optimal alignment on the sequence/pattern pair connected by the node, using an extended dynamic programming algorithm, and then 2) constructing a single common pattern from this alignment using a nested hierarchy of amino acid classes to identify the minimal inclusive amino acid class covering each paired set of elements in the alignment. Gaps within an alignment are created and/or extended using a new "pay once" gap penalty rule and gapped positions are converted into gap characters that function as 0 or 1 amino acid of any type during subsequent alignment. This method has been used to generate a library of covering patterns for homologous families in the NBRF/PIR protein sequence database. We show that a covering pattern can be more diagnostic for sequence family membership than any of the individual sequences used to construct the pattern.

---

[*] Electronic mail address: rsmith%mbcrr@harvard.edu (ARPANET); rsmith%mbcrr@husc6.bitnet (BITNET)