patterns just as easily as internally unique patterns. This is demonstrated by detection of a short internally repeated pattern in tRNA synthetases. A 3-residue pattern search was performed on 36 tRNA synthetases, including the 19 tRNA synthetases mentioned in Sections 4.2 and 4.3 along with 17 other synthetases which were removed prior to that analysis using the blast( ) operation. The short 2-residue pattern "C. .C" was found in 19 segments from 11 proteins with a significance score of $10^{-4.8}$ (segment length = 12; *minblock* = 5). This pattern is tandemly repeated in seven of these proteins; these tandem repeats were previously reported by Berg (67) as potential metal-binding domains.

4.8. DETECTING PATTERNS NEAR AMINO-TERMINAL REGIONS

All subsequences corresponding to the amino-terminal regions of every protein in the PIR protein database were selected for analysis. Low entropy sequences and all but one copy of similar sequences, related by a score of 75 or more using a PAM 120 matrix (7, 8), were eliminated. A 2-residue search revealed three significant patterns (Figure 4.12).

PATTERN	OBS	EXP	log10(prob) PTRN	log10(prob) ADJST
EF.....	44	4	-28.2	-22.7
.KK.....	84	40	-8.9	-3.4
.....L....L.....	153	92	-8.3	-2.8

```
+----+----+----+----+----+----+----+----
0    5    10   15   20   25   30   35
```

Figure 4.12. Detection of 2-residue patterns in the amino-terminal region of proteins.

Two of these patterns correspond to a pair of positively charged lysine residues (KK) and a pair of leucine residues (LL) found in the signal sequences of a number of proteins. A third pattern "EF..." with an adjusted probability < 10^{-22} was detected exactly at the amino-terminus in a large number of the proteins. Examination of these proteins revealed that most of them were fragments (i.e., partial sequences); such protein fragments typically correspond to the ends of cloned DNA fragments. Therefore, it seems likely that the

"EF..." pattern corresponds to an EcoRI restriction enzyme site, because EcoRI is commonly used in cloning experiments and recognizes the pattern "GAATTC" corresponding to the codons for glutamate (GAA) and phenylalanine (TTC). In any case, these results (and those of the previous section) reveal that ASSET is able to detect very short but significant patterns which would be missed by other methods.

5. CONCLUSION AND FUTURE WORK

5.1. ADVANTAGES AND DISADVANTAGES OF THE ASSET METHOD

The method here described offers several significant improvements over other sequence analysis methods. [1] It has a rigorous statistical basis for determining which patterns are significant and offers a quantitative measure of that significance. [2] It can efficiently search for all patterns of a given length and therefore will not miss patterns as an heuristic method might. [3] It does not require an alignment and can find patterns in a large number of sequences. [4] It can detect very short but significant patterns. [5] It can be easily modified to search for patterns having more than one residue at a specific position, which thereby presents opportunities to test for correlations between residues. [6] It is sensitive enough to search large groups of sequences not known to be related. [7] It is formulated using a mathematical model amenable to additional enhancements. Table 5.1 shows how the ASSET method compares with the method implemented in the MOTIF program (46).

Table 5.1. A comparison of MOTIF and ASSET.

FEATURE	MOTIF	ASSET
Finds patterns without alignment?	YES	YES
Finds patterns present in a majority of the proteins?	YES	YES
Finds patterns present in a minority of the proteins?	not sure	YES
Can find internally repeated patterns?	not sure	YES
Algorithm has a statistical basis?	NO	YES
Measures statistical significance?	NO	YES
Can find patterns with either of two amino acids at a single position and tests for correlation?	NO	YES
Can find patterns in proteins not known to be related?	probably not	YES

Nevertheless, there are several disadvantages of the ASSET method. [1] It does not allow for detection of sequences that have high relatedness scores but that don't exactly

match a specific pattern. [2] It cannot efficiently detect very long patterns. (This may not pose a serious problem, because a very long pattern can usually be decomposed into many short detectable patterns.) [3] It lacks a way to group together the many related patterns detected. [4] If more than one pattern is found, the method does not take into account the order with which different patterns occur in the sequences. For example, if a zinc-finger motif, and ATP/GTP-binding motifs A and B occurred in the same order in six different proteins this would be more significant then if the order of occurrence was more or less random.

5.2. FURTHER ENHANCEMENTS

Ways to overcome some of the deficits mentioned in the previous section are addressed in the next two sections.

5.2.1. INCORPORATION OF A RELATEDNESS SCORING MATRIX

The sensitivity of the ASSET method could be further increased by incorporating a relatedness scoring scheme as described in Section 2.1.1. In this modified method, a sequence would be said to match a simple pattern if a sequence-pattern relatedness score is above a specified cutoff. This score would be obtained from a PAM relatedness odds matrix (7, 8) as described in the legend to Table 5.2.

This modification could be added to the ASSET method without compromising either its statistical rigor or its algebraic system for blocks. First, note that a PAM matrix contains only integer values. Thus, for an arbitrary pattern Q, there exists some segment with a minimum (or a maximum) matching integer score, designated by $s_{\min}$ (or $s_{\max}$). Let $B_{Q,s}$ represent a block or set of segments in a population that have a relatedness score of $s_{\min} \leq s \leq s_{\max}$ to a simple pattern Q. Then, an arbitrary block $B_{Q,s}$ can be derived from elementary blocks by recursively applying the following formula starting with the universal blocks $B_{U,s} = \emptyset$ for $s \neq 0$ and $B_{U,0} = B_U$:

$$B_{Q',s'} = \bigcup_{\substack{r \in \Sigma \wedge \\ s' = s + \mathrm{PAM}[Q'(i)][r]}} \left(B_{Q,s} \cap b[r][i] \right)$$

where patterns Q and Q' differ only at position i and where $Q(i) = \Sigma$ and $|Q'(i)| = 1$. Starting from this basic formula a method could be developed for finding statistically significant patterns given a specific relatedness cutoff score.

Table 5.2. A PAM 40 matrix for calculating relatedness scores. The relatedness scores for a simple pattern Q and a subsequence S is obtained by summing the values in row $Q(i)$ and column $S(i)$ for $1 \leq i \leq |S|$.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
{A}	6	-6	-3	-3	-6	-3	-2	-1	-6	-4	-5	-6	-4	-7	-1	0	0	-12	-7	-2
{R}	-6	8	-5	-9	-7	-1	-8	-8	-1	-5	-8	1	-3	-8	-3	-2	-5	-1	-9	-7
{N}	-3	-5	7	2	-9	-3	-1	-2	1	-4	-6	0	-7	-8	-5	0	-1	-7	-4	-7
{D}	-3	-9	2	7	-12	-2	3	-3	-3	-6	-11	-4	-9	-13	-7	-3	-4	-13	-10	-7
{C}	-6	-7	-9	-12	9	-12	-12	-8	-7	-5	-13	-12	-12	-11	-7	-2	-7	-14	-3	-5
{Q}	-3	-1	-3	-2	-12	8	2	-6	1	-7	-4	-2	-3	-11	-2	-4	-5	-11	-10	-6
{E}	-2	-8	-1	3	-12	2	7	-3	-4	-5	-8	-4	-6	-12	-5	-4	-5	-15	-8	-6
{G}	-1	-8	-2	-3	-8	-6	-3	6	-8	-9	-9	-6	-7	-8	-5	-1	-5	-13	-12	-5
{H}	-6	-1	1	-3	-7	1	-4	-8	9	-8	-5	-5	-9	-5	-3	-5	-6	-6	-3	-6
{I}	-4	-5	-4	-6	-5	-7	-5	-9	-8	8	-1	-5	0	-2	-7	-6	-2	-12	-5	2
{L}	-5	-8	-6	-11	-13	-4	-8	-9	-5	-1	7	-7	1	-2	-6	-7	-6	-5	-6	-2
{K}	-6	1	0	-4	-12	-2	-4	-6	-5	-5	-7	6	-1	-12	-6	-3	-2	-10	-8	-8
{M}	-4	-3	-7	-9	-12	-3	-6	-7	-9	0	1	-1	11	-3	-7	-5	-3	-11	-10	-1
{F}	-7	-8	-8	-13	-11	-11	-12	-8	-5	-2	-2	-12	-3	9	-9	-6	-8	-4	2	-7
{P}	-1	-3	-5	-7	-7	-2	-5	-5	-3	-7	-6	-6	-7	-9	8	-1	-3	-12	-12	-5
{S}	0	-2	0	-3	-2	-4	-4	-1	-5	-6	-7	-3	-5	-6	-1	6	1	-4	-6	-5
{T}	0	-5	-1	-4	-7	-5	-5	-5	-6	-2	-6	-2	-3	-8	-3	1	7	-11	-6	-2
{W}	-12	-1	-7	-13	-14	-11	-15	-13	-6	-12	-5	-10	-11	-4	-12	-4	-11	13	-4	-14
{Y}	-7	-9	-4	-10	-3	-10	-8	-12	-3	-5	-6	-8	-10	2	-12	-6	-6	-4	10	-6
{V}	-2	-7	-7	-7	-5	-6	-6	-5	-6	2	-2	-8	-1	-7	-5	-5	-2	-14	-6	7
Σ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

5.2.2. A Method for Combining Related Patterns into Groups

The ASSET method often detects a large number of patterns; in order to facilitate the interpretation of these patterns by revealing relationships between them, it would be helpful to have a procedure that arranges them into a hierarchy of subpatterns and superpatterns. For example, Figure 5.1 shows a group of related patterns detected by ASSET in 29 reverse transcriptases. These patterns were aligned by hand and are subpatterns of the pattern, "G.PYNPQ.QG.VER".

			log10	
PATTERN	OBS	EXP	PROB	SIGNIF
....YN......VE........	9	0.13	-13.5	-5.2
...P........VER.......	10	0.23	-13.1	-4.8
...P.N......VE........	10	0.23	-13.1	-4.7
...PY.......VE........	9	0.15	-12.9	-4.6
....Y.P.....VE........	9	0.15	-12.9	-4.6
.G.P........VE........	10	0.25	-12.7	-4.3
.G.....Q.Q....R.......	9	0.17	-12.5	-4.1
.G...N.Q.Q............	9	0.18	-12.4	-4.1
.G.....Q.Q..V.........	9	0.19	-12.2	-3.8
.G...N.Q....V.........	9	0.19	-12.2	-3.8
.G.P...Q.Q............	9	0.20	-11.8	-3.5
...P.N.......ER.......	9	0.21	-11.7	-3.4
....Y.......VER.......	8	0.13	-11.7	-3.4
....Y..Q....VE........	8	0.13	-11.7	-3.3
.G.......Q..VE........	9	0.22	-11.6	-3.3
.G.P...Q....V.........	9	0.22	-11.6	-3.3
.G...N......VE........	9	0.22	-11.6	-3.3
.G.P.....Q..V.........	9	0.22	-11.6	-3.3
......P.....VER.......	9	0.22	-11.5	-3.1
...PYN.......E........	8	0.14	-11.4	-3.1
....Y.....G.VE........	8	0.14	-11.4	-3.0
....Y.P...G.V.........	8	0.14	-11.4	-3.0
...P.....Q..VE........	9	0.23	-11.4	-3.0

.G.PYNPQ.QG.VER.......				

Figure 5.1. A group of 23 related patterns detected by ASSET in 29 reverse transcriptases. A superpattern for the group is given at the bottom of the figure.

5.3. TOOL DEVELOPMENT AND FUTURE APPLICATIONS

5.3.1. TOOLS FOR RETRIEVING GROUPS OF PROTEINS

The utility of the ASSET method could be extended by developing tools for retrieving groups of functionally or pattern related sequences from the databases. For example, if a significant pattern is detected in a small group of proteins, it would be helpful to have a tool for retrieving every protein from the database that also contains that pattern. As a second example, in order to detect motifs characteristic of a specific cofactor binding site, it would be helpful to have a tool for retrieving all proteins binding to that cofactor.

5.3.2. SEED PATTERN BLOCKS

In Section 4.6 a procedure was described for detecting distantly related proteins using a seed pattern block. However, in order to exhaustively search a protein database by repeatedly using this procedure, a large number of seed pattern blocks would be needed. Purging closely related segments from these blocks would be a formidable task, but could be accomplished through parallel processing.

5.3.3. APPLICATION TO NUCLEIC ACID SEQUENCES

The ASSET method can be applied to the analysis of nucleic acids simply by redefining the alphabet. Examples of biological sequence entities that might be chosen for analysis are origins of replication, regions where precursor mRNA splicing or transposon insertion occurs, or regulatory protein binding sites. Segments could be generated from these sequences and analyzed for patterns common to these regions.

6. APPENDIX

6.1. DEFINITIONS OF TERMS

Table 6.1. Definitions and descriptions of terms used in the text.

Symbol	Description	Definition
N	The set of all natural numbers	{0,1,2,3,4,...}
P	The set of all positive integers	{1,2,3,4,...}
I	The set of all intergers	{...,-3,-2,-1,0,+1,+2,+3,...}
R	The set of all real numbers	–
Σ	amino acid or nucleotide alphabet	{A,C,D,E,F,G,H,I,K,L,M,N,P,Q, R,S,T,V,W,Y} or {A,C,G,T}
$\mathcal{P}(\Sigma)$	The power set of Σ	$\{ s \mid s \subseteq \Sigma\}$
Σ^*	The set of all strings of elements of Σ	$\Sigma^* = \{ \langle r_1,r_2, ...,r_n\rangle \mid r_i \in \Sigma \wedge 1 \le i \le n \wedge n \in \mathbf{N} \}$
Σ^+	The set of all non-**null** strings of elements of Σ	$\Sigma^+ = \{ \langle r_1,r_2, ...,r_n\rangle \mid r_i \in \Sigma \wedge 1 \le i \le n \wedge n \in \mathbf{P} \}$
S	A biological sequence	$S \in \Sigma^+$
I	An entity identifier	$I \in \mathbf{N}$
E	A biological sequence entity	$E = \langle I,S\rangle \in F_{Seq}$: $\mathbf{N} \rightarrow \Sigma^+$ such that $E.S = F_{Seq}(E.I)$
Q	A pattern	$Q \in \{\langle s_1,...,s_k\rangle \mid s_i \in \mathcal{P}(\Sigma) \wedge k \in \mathbf{P}\}$
Q_U	The universal pattern	$Q_U \equiv \langle s_1,...,s_k\rangle : s_i = \Sigma \wedge 1 \le i \le k$
q	A simple pattern	$q \in \{\langle s_1,...,s_k\rangle \mid s_i \in \{\Sigma\} \cup \{\{r\} \mid r \in \Sigma\} \wedge 1 \le i \le k \}$
'.'	The set containing Σ	'.' $\equiv \{\Sigma\}$
'?'	The set of all singleton sets	'?' $\equiv \{\{x\} \mid x \in \Sigma\}$
G	The set of all pattern generators	$\mathbf{G} \equiv \{\langle \sigma_1,...,\sigma_k\rangle \mid \sigma_i \subseteq \mathcal{P}(\Sigma) \wedge k \in \mathbf{P}\}$
G	A pattern generator	$G \in \mathbf{G}$
$G_\emptyset$	The null generator	a sequence of 0 sets
G_Q	A 1-pattern generator	$G \in \mathbf{G}$: G only matches pattern Q
G_v	A variable residue generator	G_v is derived from a 1-pattern generator, G_Q, by converting two '.'-sets in G_Q into '?'-sets.

Table 6.1 (continued): Definitions and descriptions of terms used in the text.

Symbol	Description	Definition
G_t	A 2×2 table generator	G_t is derived from a 1-pattern generator, G_Q, by 2-partitioning two sets in Q.
e	A sequence segment	$e = \langle I: \mathbf{N}, o: \mathbf{N}, k: \mathbf{P} \rangle$ where $\langle e, S \rangle \in f_{seg}$
f_{seg}	The segment function	$f_{seg}(e) = r_{e.o+1} r_{e.o+2} \ldots r_{e.o+e.k} \in \Sigma^*$ where $\exists E : E.S = F_{Seq}(e.I) = r_1 r_2 \ldots r_n \wedge 0 \leq e.o \leq n - e.k.$
Π	An arbitrary set of biological sequence entities	$\Pi \subseteq F_{Seq}$
β_k	The set of all length k segments derivable from Π	$\beta_k \equiv \{ e \mid \langle e, S \rangle \in f_{seg} \wedge e.k = k \wedge e.I = E.I \wedge E \in \Pi \}$
$b[r][i]$	An elementary block	$b[r][i] \equiv \{ e \in \beta_k \mid S = f_{seg}(e) \wedge S(i) = r \}$ where $r \in \Sigma \wedge i \in \mathbf{N}$
B_U	The universal block	$B_U \subseteq \beta_k$
$B_\emptyset$	The null block	$\emptyset$
B_Q	A pattern block	$B_Q = \{ e \in B_U \mid f_{seg}(e)$ matches $Q \}$
$\mathbf{P_{seg}}$	The set of segment populations	$\mathbf{P_{seg}} \equiv \{ \langle \Pi, k, B_U, \Sigma \rangle \mid B_U \subseteq \beta_k \}$
P	An aligned segment population	$P \in \mathbf{P_{seg}}$
$P_\emptyset$	The null population	$P_\emptyset = \langle \emptyset, 0, \emptyset, \Sigma \rangle$
D	A generator-population dyad	$D \in \{ \langle P, G \rangle \mid P \in \mathbf{P_{seg}} \wedge G \in \mathbf{G} \}$

6.2. STATISTICAL FORMULAS

This section describes the statistical functions used in the ASSET method.

6.2.1. CUMULATIVE BINOMIAL DISTRIBUTION

The cumulative binomial distribution function is related to the incomplete beta function $I_x(a, b)$ as follows (68):

$$\sum_{i=k}^{N} \binom{N}{i} p^i (1-p)^{N-i} = I_p(k, N - k + 1) \tag{6.1}$$

Therefore, pattern probabilities were determined using an efficient algorithm for the incomplete beta function (68).

6.2.2. FISHER'S EXACT TEST

A significance test for independence in a 2x2 table is performed using Fisher's exact test (60), which can be used no matter how small the numbers in the table happen to be. The significance is obtained by calculating the multinomial probability of obtaining the observed table, under the assumption of independence and conditional on the marginal totals being what they are, together with the probabilities of obtaining all tables, having the same marginal totals, which are more extreme (i.e., have a lower multinomial probability) than the observed table. Therefore, given the following table,

	column 1	column 2	total
row 1	a	b	$a+b$
row 2	c	d	$c+d$
total	$a+c$	$b+d$	n

the probability of obtaining the observed numbers, conditional on the marginal totals is given by,

$$\frac{\binom{n}{a\ b\ c\ d}}{\binom{n}{a+b}\binom{n}{b+d}}$$

$$= \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{(n!a!b!c!d!)}$$

where $n = a + b + c + d$. Therefore, the exact test can be performed using Equation 6.2, which is the summation of the probabilities for all tables with the same marginal totals as the observed table but with equivalent or smaller probabilities (this is a two sided test).

$$\sum_{i=\max(-b,-c)}^{\min(a,d)} [a!b!c!d! \le (a-i)!(b+i)!(c+i)!(d-i)!] \; \frac{(a+b)!(a+c)!(b+d)!(c+d)!}{n!(a-i)!(b+i)!(d-i)!(c+i)!} \qquad (6.2)$$

The expression enclosed in square brackets (designated as $[P]$) has the value 1 if the boolean expression P is true and the value 0 if P is false.

6.2.3. "LOW ENTROPY" SEGMENTS

A mathematically rigorous description of low entropy sequences has not yet been formulated. However, as a first approximation, "low entropy" segment scores are determined using the negative logarithm of the multinomial probability function. Thus, using this scoring method, a "low entropy" sequence contains a disproportionate number of certain residues, relative to the frequencies of those residues in the population as a whole. Therefore, such a sequence has a low multinomial probability P, which is obtained from the following equation,

$$P = \binom{n}{n_1 \;\; n_2 \;\; \ldots \;\; n_{|\Sigma|}} p_1^{n_1} p_2^{n_2} \ldots p_{|\Sigma|}^{n_{|\Sigma|}} \qquad (6.3)$$

where n is the length of the segment, n_i is the number of r_i residues in the segment, p_i is the frequency of r_i in the population as a whole, and $r_i \in \Sigma$, $n_i \in \mathbf{N}$, $p_i \in \mathbf{R}$, $\sum_i p_i = 1.0$ and $1 \le i \le |\Sigma|$. The multinomial coefficient in Equation 6.3 can be obtained from a series of binomial coefficients, using the equality:

$$\binom{n}{n_1 \;\; n_2 \;\; \ldots \;\; n_{|\Sigma|}} = \binom{n}{n_1} \binom{n-n_1}{n_2} \cdots \binom{n-n_1-n_2-\ldots-n_{|\Sigma|-2}}{n_{|\Sigma|-1}} \qquad (6.4)$$

Thus "low entropy" scores are determined using an efficient algorithm for the binomial coefficient (68) and Equation 6.4 during the purgeH( ) operation. (The negative logarithm of the multinomial probability is used only to simplify the calculations.) A low entropy cutoff score is obtained by taking the average of the highest entropy scores for several

Monte Carlo simulations (this gives a rough estimate of the highest score expected by chance).

6.3. TABLE ABSTRACT DATA TYPE

The table abstract data type is used to create contingency tables and to perform Fisher's exact test for significant correlations between residues at different positions in a pattern. This data type is used in conjunction with a 2x2 table generator as described in Section 3.3.

6.3.1. DEFINITION

A **table** (T), corresponding to the 2x2 contingency table shown in Figure 6.1, is defined as a 2-tuple

$$T = \langle bin: (\mathcal{P}(\Sigma)^2)\text{x}(\mathcal{P}(\Sigma)^2),\ cell: (\mathbf{N}^2)\text{x}(\mathbf{N}^2)\rangle$$

where $cell$ = $\langle\langle a,b\rangle, \langle c,d\rangle\rangle$ and $bin = \langle\langle s_r, s_{r'}\rangle, \langle s_c, s_{c'}\rangle\rangle$.

	s_c	$s_{c'}$	
s_r	a	b	$a+b$
$s_{r'}$	c	d	$c+d$
	$a+c$	$b+d$	$a+b+c+d$

Figure 6.1. A 2x2 contingency table.

A **null table**, $T\emptyset$, is a table with all cells initialized to zeros, $T\emptyset.cell$ = <<0,0>, <0,0>>, and all bins undefined, $T\emptyset.bin$ = <<**null**,**null**>, <**null**,**null**>>.

6.3.2. OPERATIONS

The following operation is used to create a table.

make_table( ): Create and return a **null** table $T\emptyset$.

Two operation are used to assign values to the bins and cells of a table.

assign_bins(**integer** i, **partition** $\{s_0,s_1\}$, **table** T): If $i \in \{0,1\}$ and $s_0 \cap s_1 = \emptyset$ and $s_0, s_1 \neq \emptyset$, then assign $T.bin(i)(0) := s_0$ and $T.bin(i)(1) := s_1$ and return T; otherwise destroy T and return **null**.

assign_cell(**integer** *i*, **integer** *j*, **integer** *n*, **table** *T*): If $i,j \in \{0,1\}$, then assign $T.cell(i)(j) := n$.

Access to the table bins, which is needed by the generator operation tab_pattern( ), is available using the following operation.

get_bin(**integer** *i*, **integer** *j*, **table** *T*): If $i,j \in \{0,1\}$, then return$T.bin(i)(j)$; otherwise return **null**.

The significance of correlation for a table is determined using Fisher's exact test.

exact_test(**table** *T*): Calculate the probability of correlation using Fisher's exact test on table *T*.

6.4. PROOF FOR $P_{ADJ} < NP + (NP)^2$

For the following proof, it is assumed that $N,i \in \mathbf{P}$, $P_{adj},p,k \in \mathbf{R}$, $k \geq 1$ and $p = \frac{1}{kN}$.

Lemma 1:

$$\binom{N}{i+1}p^{i+1} < \frac{1}{2k}\binom{N}{i}p^{i} \quad \text{where } 1 \leq i < N.$$

Proof:

$p = \frac{2}{2kN}$	by assumption
$\rightarrow p \leq \frac{i+1}{2kN}$	because $i \geq 1$
$\rightarrow p < \frac{i+1}{2k(N-i)}$	because $\frac{1}{N} < \frac{1}{N-i}$
$\rightarrow N!p^{i+1} < \frac{i+1}{2k(N-i)}N!p^{i}$	multiply both sides by $N!p^i$
$\rightarrow \frac{N!}{i!(N-i-1)!}p^{i+1} < \frac{i+1}{2k(N-i)i!(N-i-1)!}N!p^{i}$	multiply both sides by $\frac{1}{i!(N-i-1)!}$
$\rightarrow \frac{N!}{(i+1)!(N-(i+1))!}p^{i+1} < \frac{1}{2k}\frac{N!}{i!(N-i)!}p^{i}$	rearrange
$\rightarrow \binom{N}{i+1}p^{i+1} < \frac{1}{2k}\binom{N}{i}p^{i}$	QED.

Theorem 6.1:

$$P_{adj} = \sum_{i=1}^{N} \binom{N}{i} p^i (1\text{-}p)^{N-i} \rightarrow P_{adj} < Np + (Np)^2$$

Proof:

$$P_{adj} = \sum_{i=1}^{N} \binom{N}{i} p^i (1\text{-}p)^{N-i} \qquad \text{by assumption}$$

$$< \sum_{i=1}^{N} \binom{N}{i} p^i \qquad \text{because } (1\text{-}p)^{N-i} < 1$$

$$= \binom{N}{1} p^1 + \binom{N}{2} p^2 + \ldots + \binom{N}{N\text{-}1} p^{N-1} + \binom{N}{N} p^N \qquad \text{by definition}$$

$$< \binom{N}{1} p^1 + \binom{N}{2} p^2 + \ldots + \left(1 + \frac{1}{2k}\right)\binom{N}{N\text{-}1} p^{N-1} \qquad \text{by lemma 1}$$

$$< \binom{N}{1} p^1 + \ldots + \left(1 + \frac{1}{2k}\left(1 + \frac{1}{2k}\right)\right)\binom{N}{N\text{-}2} p^{N-2} \qquad " \quad "$$

$$= \binom{N}{1} p^1 + \ldots + \left(1 + \frac{1}{(2k)^1} + \frac{1}{(2k)^2}\right)\binom{N}{N\text{-}2} p^{N-2} \qquad \text{rearrange}$$

$$\vdots$$

$$< \left(1 + \frac{1}{(2k)^1} + \frac{1}{(2k)^2} + \ldots + \frac{1}{(2k)^{N-1}}\right)\binom{N}{1} p^1 \qquad \text{by lemma 1}$$

$$< \left(\sum_{j \geq 0} \frac{1}{(2k)^j}\right) Np$$

$$= \left(\frac{1}{1 - \frac{1}{2k}}\right) Np \qquad \text{sum of a geometric series}$$

$$= \left(\frac{2k}{2k - 1}\right) Np = \left(\frac{1 + 2k - 1}{2k - 1}\right) Np \qquad \text{rearrange}$$

$$= \left(1 + \frac{1}{2k-1}\right)Np \leq \left(1 + \frac{1}{k}\right)Np \qquad \text{since } k \geq 1$$

$$= (1 + Np)Np = Np + (Np)^2 \qquad \text{since } k = \frac{1}{pN}, \text{ QED.}$$

6.5. PROTEIN SEQUENCE ENTITIES

The protein sequences used in this study are as follows. The SWISS-PROT (69) primary accession numbers for the 19 tRNA synthetases are P21888, P22438, P07814, P14325, P11875, P23395, P00958, P13188, P07806, P23381, P04803, P22249, P09436, P00956, P10857, P15181, P00959, P12063, P00951. 29 proteins were extracted from the PIR (version 31) database having a 30 residue segment exactly matching 29 of the 33 reverse transcriptases given by Smith et al. (46) (no databank accession numbers were given by Smith et al.); four of the 33 proteins, designated by Smith et al. as TYS2, IFAC, C2IS, and INGT, were not found. The PIR ENTRY identification codes for the 29 proteins are: ALRCG_2, BLVGAGA_2, CANRVDNA_5, DROELEF_2, DROGYPSY_1, DROVLPHR_1, HERGPE_1, HIVHTLV3_2, HIVV2RODX_4, HPBADRA_1, HTVPROP_2, HUMHIVPOL_1, HUMTNL12_2, MCAMVG1_5, MLFFMLVCGD_2, MMTPROCG_2, RATL1RTO2B_1, SIVMPCG_2, VLVCGA_2, YSCTYA117_2, GNHUER, GNHYIH, GNLJEW, GNFF17, GNFF42, QXBY31, A25657, C24785, and B24872. Similarly, 15 proteins were extracted from the PIR (version 31) and SWISS-PROT (version 21) databases having a 30 residue segment exactly matching the 15 adenine methylases given by Smith et al. (46). The PIR ENTRY identification codes for 13 of the proteins are: CHV1AMB3_1, ECOP15BMO_1, FVBFOKMR_1, HEAHHAIIMT_1, HEAMTEN_1, PLBECORV_2, PP1MOD_1, PROIRM_1, PSEPAER7_1, PT4T4G69_2, RI1ECOR_2, STRDPN2A_1, STRDPN2A_2; the SWISS-PROT primary accession numbers for the 2 remaining proteins are P00475, and P14385.

6.6. SAMPLE SESSION

Below is a sample ASSET session demonstrating several of the commands described in Chapter 3. The input file contained a set of 15 methyltransferases.

```
neuwald@wuibc3 69> nasset methyltransferases
  5840/5840 segments from 15/15 entities
  segment length: 15
asset> reset 12
asset> search 3
        input file: 'methyltransferases'
        5840 segments from 15 entities
        segment length: 12
        minimum search block size: 3
        minimum output block size: 5
        search depth: 3
        heap size: 500
        print if adjusted probability < 0.01 (1e-2.0)

Saved 500 out of 444400 patterns
                                          log10(prob)
PATTERN        OBS   (E)      EXP      PTRN     ADJST
PPY.........    14( 13)     0.49    -15.5     -9.9
D.....PP....    11( 10)     0.59    -10.3     -4.7
DP.Y........    11( 10)     0.69     -9.7     -4.0
DPP.........    10(  9)      0.6     -9.0     -3.4
D......PY...    10(  9)     0.68     -8.5     -2.9
D.....P.Y...    10(  9)     0.68     -8.5     -2.9
D.PY........    10(  9)     0.69     -8.5     -2.8
+----+----+-
0    5    10    7 hits

Merged patterns:
                                          log10(prob)
PATTERN        OBS   (E)      EXP      PTRN     ADJST
N.PY........    16( 13)      1.4    -11.5     -3.7
D
NPP.........    15( 14)      1.2    -11.2     -3.4
D
D.....P.Q...    14( 11)      1.1    -11.0     -3.3
         Y
Y..PP.......    13( 11)      1.1     -9.8     -2.1
V
+----+----+-
0    5    10    4 merged patterns
        time: 57 seconds (0.95 minutes)

asset> ?..PPY
                                          log10(prob)
PATTERN        OBS   (E)      EXP      PTRN     ADJST
I..PPY......     1(  1)    0.032     -1.5      --
L..PPY......     1(  1)    0.047     -1.3      --
F..PPY......     2(  2)    0.027     -3.5      --
Y..PPY......     8(  7)    0.023    -17.7      --
V..PPY......     2(  2)    0.028     -3.4      --
```

```
asset> {ILV:YF}.{N:D}PPY

  (0,2)    N          D          PROB.    DEPEND
  ILV [ 4(1.1)     0(2.9) ]     0.001    1.0
   YF [ 0(2.9)    10(7.1) ] (total:14)

asset> {ILVYF}.{ND}PPY
                                        _log10(prob)_
PATTERN        OBS   (E)     EXP      PTRN     ADJST
I.NPPY......    14( 13)   0.019    -35.0      --
L D
F
Y
V

asset> reset 20
asset> block ....{ILVYF}.{ND}PPY

SEQUENCE                      IDENTITY          LOCATION
fdfi V g NPPY vvrpsgyknd   CHV1AMB3_1       (112-131)
vkmi Y i DPPY ntgkdgfvyn   ECOP15BMO_1     (116-135)
gdil Y i DPPY ngrqyisnyh   FVBFOKMR_1      (211-230)
ndlv Y c DPPY littgsyndg   FVBFOKMR_1      (541-560)
idli F a DPPY fmqtegkllr   HEAMTEN_1       (29-48)
ddvv Y c DPPY igrhvdyfns   PLBECORV_2      (186-205)
vnmi Y i DPPY ntgkdgfvyn   PP1MOD_1        (116-135)
ynka I l NPPY lkiaakgrer   PROIRM_1        (146-165)
fdfv V g NPPY vrpelipapl   PSEPAER7_1      (113-132)
gdfv Y v DPPY litvadynkf   PT4T4G69_2      (164-183)
gdfv Y f DPPY iplsetsaft   STRDPN2A_1      (187-206)
mdmi F a DPPY flsnggisns   STRDPN2A_2      (29-48)
asvv Y c DPPY aplsatanft   DMA_ECOLI       (174-193)
fdli L g NPPY givgeaskyp   MTTA_THEAQ      (98-117)
_____________________
.... I . NPPY .........    14 segments from 13 entities
     L   D
     F
     Y
     V

asset> select .....P
   239/5720 segments selected
asset> search 3
        input file: 'methyltransferases'
        239 segments from 15 entities
        segment length: 20
        minimum search block size: 3
        minimum output block size: 5
        search depth: 3
        heap size: 500
        print if adjusted probability < 0.01 (1e-2.0)

Saved 500 out of 7.8204e+06 patterns
                                        _log10(prob)_
PATTERN              OBS   (E)     EXP     PTRN     ADJST
....DPPY.........    10(  9)    0.13    -15.5     -8.6
...DPPY..........    10(  9)    0.16    -14.6     -7.7
```

```
.Y..PPY.............     8(  7)    0.11   -12.4    -5.5
.Y.D.PY.............     8(  7)    0.12   -12.2    -5.3
..Y.DP.Y............     8(  7)    0.12   -12.1    -5.2
.Y.DPP..............     8(  7)    0.12   -12.0    -5.1
..Y.DPP.............     8(  7)    0.13   -11.9    -5.0
..Y..PPY............     8(  7)    0.14   -11.5    -4.7
+----+----+----+----
0    5    10   15       8 hits

Merged patterns:
                                                _log10(prob)_
PATTERN                OBS   (E)      EXP      PTRN     ADJST
....NPPY.............   14( 13)     0.26    -19.3    -10.9
    D
...NPPY..............   14( 13)     0.32    -18.2     -9.8
   D
I...PPY..............   12( 12)     0.26    -16.0     -7.6
V
.I...PPY.............   12( 12)     0.32    -14.8     -6.4
 V
I..D.PY..............    9(  9)     0.27    -10.9     -2.5
V
.I..DP.Y.............    9(  9)     0.28    -10.7     -2.3
 V
M....PPY.............    7(  7)      0.1    -10.7     -2.3
F
I..DPP...............    9(  9)     0.28    -10.7     -2.3
V
.I..DPP..............    9(  9)     0.29    -10.6     -2.2
 V
+----+----+----+----     9 merged patterns
0    5    10   15
        time: 27 seconds (0.45 minutes)

asset> reset 12
asset> blast 150
15 items compared
13 equiv. classes
4961/5840 segments selected from 13/15 entities

asset> revert
asset> shuffle
asset> search 3
        input file: (shuffled_sequences)
        5840 segments from 15 entities
        segment length: 12
        minimum search block size: 3
        minimum output block size: 5
        search depth: 3
        heap size: 500
        print if adjusted probability < 0.01 (1e-2.0)

Saved 500 out of 444400 patterns
        0 hits
        time: 57 seconds (0.95 minutes)
asset> quit
neuwald@wuibc3 35>
```

7. BIBLIOGRAPHY

1. G.N Reeke, Jr, "Protein folding: computational approaches to an exponential-time problem," *Annual Review of Computer Science*, vol. 3, pp. 59-84, 1988.

2. G.E. Schulz, "A critical evaluation of methods for predictiion of protein secondary structures," *Annual Review of Biophysics and Biophysical Chemistry*, vol. 17, pp. 1-21, 1988.

3. M. Saraste, P.R. Sibbald, and A. Wittinghofer, "The P-loop - a common motif in ATP- and GTP-binding proteins," *Trends in the Biological Sciences*, vol. 15, pp. 430-434, 1990.

4. A.F. Neuwald, J.D. York, and P.W. Majerus, "Diverse proteins homologous to inositol monophosphatase," *FEBS Letters*, vol. 294, pp. 16-18, 1991.

5. J.B. Kruskal, "An overview of sequence comparison," In *Time Warps, String Edits, and Macromolecules: the Theory and Practice of Sequence Comparison*, D. Sankoff, and J.B. Kruskal, eds., pp. 1-44, Addison-Wesley, Reading, 1983.

6. D.J. Bacon, and W.F. Anderson, "Multiple sequence alignment," *Journal of Molecular Biology*, vol. 191, pp. 153-161, 1986.

7. M.O. Dayhoff, R.M. Schwartz, and B.C. Orcutt, "A model of evolutionary change in proteins," In *Atlas of Protein Sequence and Structure*, M.O. Dayhoff, ed., vol. 5 suppl. 3, pp. 345-352, National Biomedical Research Foundation, Washington D.C, 1978 .

8. R.M. Schwartz, and M.O. Dayhoff, "Matrices for detecting distant relationships," In *Atlas of Protein Sequence and Structure*, M.O. Dayhoff, ed., vol. 5 suppl. 3, pp. 353-358, Nat. Biomed. Res. Found., Washington D.C, 1978.

9. W.M. Fitch, and T.F. Smith, "Optimal sequence alignments," *Proceedings Of The National Academy Of Sciences, USA*, vol. 80, pp. 1382-1386, 1983.

10. S.B. Needleman, and C.D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, pp. 443-453, 1970.

11. P.H. Sellers, "An algorithm for the difference between two finite sequences," *Journal of Combinatorial Theory*, vol. 16, pp. 253-258, 1974.

12. P.H. Sellers, "On the theory and computation of evolutionary distances," *SIAM Journal of Applied Mathematics*, vol. 26, pp. 787-793, 1974.

13. T.F. Smith, and M.S. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology*, vol. 147, pp. 195-197, 1981.

14. D.J. Lipman, and W.R. Pearson, "Rapid and sensitive protein similarity searches," *Science*, vol. 227, pp. 1435-1441, 1985.

15. W.R. Pearson, and D.J. Lipman, "Improved tools for biological sequence comparison," *Proceedings of the National Academy of Sciences USA*, vol. 85, pp. 2444-2448, 1988.

16. W.J. Wilbur, and D.J. Lipman, "Rapid similarity searches of nucleic acid and protein data banks," *Proceedings Of The National Academy Of Sciences USA*, vol. 80, pp. 726-730, 1983.

17. W.J. Wilbur, and D.J. Lipman, "The context dependent comparison of biological sequences," *SIAM Journal of Applied Mathematics*, vol. 44, pp. 557-567, 1984.

18. S.F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic Local Alignment Search Tool," *Journal of Molecular Biology*, vol. 215, pp. 403-410, 1990.

19. J.E. Hopcroft, and J.D. Ullman, *Introduction to Automata Theory, Languages, and Computation*, Addison-Wesley, Reading, MA, 1979.

20. S.F. Altschul, and D.J. Lipman, "Trees, stars, and multiple biological sequence alignment," *SIAM Journal of Applied Mathematics*, vol. 49, pp. 197-209, 1989.

21. G.J. Barton, and M.J.E. Sternberg, "A strategy for the rapid multiple alignment of protein sequences," *Journal of Molecular Biology*, vol. 198, pp. 327-337, 1987.

22. H. Carrillo, and D.J. Lipman, "The multiple sequence alignment problem in biology," *SIAM Journal of Applied Mathematics*, vol. 48, pp. 1073-1082, 1988.

23. D.-F. Feng, and R.F. Doolittle, "Progressive sequence alignment as a prerequisite to correct phylogenetic trees," *Journal of Molecular Evolution*, vol. 25, pp. 351-360, 1987.

24. M.S. Johnson, and R.F. Doolittle, "A method for the simultaneous alignment of three or more amino acid sequences," *Journal of Molecular Evolution*, vol. 23, pp. 267-278, 1986.

25. D.J. Lipman, S.F. Altschul, and J.D. Kececioglu, "A tool for multiple sequence alignment," *Proceedings of the National Academy of Sciences USA*, vol. 86, pp. 4412-4415, 1989.

26. M. Murata, J.S. Richardson, and J.L. Sussman, "Simultaneous comparison of three protein sequences," *Proceedings of the National Academy of Sciences USA*, vol. 82, pp. 3073-3077, 1985.

27. D. Sankoff, "Minimal mutation trees of sequences," *SIAM Journal of Applied Mathematics*, vol. 28, pp. 35-42, 1975.

28. D. Sankoff, and R.J. Cedergren, "Simultaneous comparison of three or more sequences related by a tree," In *Time Warps, String Edits, and Macromolecules: the Theory and Practice of Sequence Comparison*, D. Sankoff, and J.B. Kruskal, eds., pp. 253-263, Addison-Wesley, Reading, 1983.

29. W.R. Taylor, "Multiple sequence alignment by a pairwise algorithm," *Computer Applications in the Biosciences*, vol. 3, pp. 81-87, 1987.

30. M. Vingron, and P. Argos, "A fast and sensitive multiple sequence alignment algorithm," *Computer Applications in the Biosciences*, vol. 5, pp. 115-121, 1989.

31. M.S. Waterman, and M. Perlwitz, "Line geometries for sequence comparisons," *Bulletin of Mathematical Biology*, vol. 46, pp. 567-577, 1984.

32. G.D. Stormo, and G.W. Hartzell, " Identifying protein-binding sites from unaligned DNA fragments," *Proceedings Of The National Academy Of Sciences USA*, vol. 86, pp. 1183-1187, 1989.

33. E. Sobel, and H.M. Martinez, "A multiple sequence alignment program," *Nucleic Acids Research*, vol. 14, pp. 363-374, 1986.

34. T.H. Byers, and M.S. Waterman, "Determining all optimal and near-optimal solutions when solving shortest path problems by dynamic programming," *Operations Research*, vol. 46, pp. 473, 1984.

35. S. Karlin, M. Morris, G. Ghandour, and M.-Y. Leung, "Efficient algorithms for molecular sequence analysis," *Proceedings of the National Academy of Sciences USA*, vol. 85, pp. 841-845, 1988.

36. G.D. Schuler, S.F. Altschul, and D.J. Lipman, "A workbench for multiple alignment construction and analysis," *Proteins*, vol. 9, pp. 180-191, 1991.

37. S. Karlin, and S.F. Altschul, "Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes," *Proceedings Of The National Academy Of Sciences USA*, vol. 87, pp. 2264-2268, 1990.

38. S. Karlin, A. Dembo, and T. Kawabata "Statistical composition of high-scoring segments from molecular sequences," *Annals of Statistics*, vol. 18, pp. 571-581, 1990.

39. M. Vingron, and P. Argos, "Motif recognition and alignment for many sequences by comparison of dot-matrices," *Journal of Molecular Biology*, vol. 218, pp. 33-43, 1991.

40. M. Gribskov, R. Luethy, and D. Eisenberg, "Profile analysis," *Methods in Enzymology*, vol. 183, pp. 146-159, 1990.

41. M. Gribskov, M. McLachlan, and D. Eisenberg, "Profile analysis: detection of distantly related proteins," *Proceedings of the National Academy of Sciences USA*, vol. 84, pp. 4355-4358, 1987.

42. M. Gribskov, M. Homyak, J. Edenfield, and D. Eisenberg, "Profile scanning for three-dimensional structural patterns in protein sequences," *Computer Applications in the Biosciences*, vol. 4, pp. 61-66, 1988.

43. C.L. Queen, M.N. Wegman, and L.J. Korn, "Improvements to a program for DNA analysis: a procedure to find homologies among many sequences," *Nucleic Acids Research*, vol. 10, pp. 449-456, 1982.

44. M.S. Waterman, R. Arratia, and D.J. Galas, "Pattern recognition in several sequences: consensus and alignment," *Bulletin of Mathematical Biology*, vol. 46, pp. 515-527, 1984.

45. R. Staden, "Methods for discovering novel motifs in nucleic acid sequences," *Computer Applications in the Biosciences*, vol. 5, pp. 293-298, 1989.

46. H.O. Smith, T.M. Annau, and S. Chandrasegaran, "Finding sequence motifs in groups of functionally related proteins," *Proceedings of the National Academy of Sciences USA*, vol. 87, pp. 826-830, 1990.

47. S. Henikoff, and J.G. Henikoff, "Automatic generation of protein blocks for database searching," *Nucleic Acids Research*, vol. 19, pp. 6565-6572, 1991.

48. C.G. Davis, "The many faces of epidermal growth factor repeats," *New Biologist*, vol. 2, pp. 410-419, 1990.

49. A. Klug, and D. Rhodes, "'Zinc fingers': a novel protein motif for nucleic acid recognition," *Trends in the Biological Sciences*, vol. 12, pp. 464-469, 1987.

50. J. Posfai, A.S. Bhagwat, G. Posfai, and R.J. Roberts, "Predictive motifs derived from cytosine methyltransferases," *Nucleic Acids Research*, vol. 17, no. 7, pp. 2421-2435, 1989.

51. C.E. Lawrence, and A.A. Reilly, "An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences," *Proteins*, vol. 7, pp. 41-51, 1990.

52. R.J.A. Little, and D.B. Rubin, *Statistical Analysis With Missing Data*, John Wiley & Son, New York, 1987.

53. K.Y. Cockwell, and I.G. Giles, "Software tools for motif and pattern scanning: program descriptions including a universal sequence reading algorithm," *Computer Applications in the Biosciences*, vol. 5, pp. 227-232, 1989.

54. M.J.E. Sternberg, "PROMOT: A FORTRAN program to scan protein sequences against a library of known motifs," *Computer Applications in the Biosciences*, vol. 7, pp. 257-260, 1991.

55. P.R. Sibbald, and P. Argos, "Scrutineer: A computer program that flexibly seeks and describes motifs and profiles in protein sequence databases," *Computer Applications in the Biosciences*, vol. 6, pp. 279-288, 1990.

56. E.W. Myers, and W. Miller, "Approximate marching of regular expressions," *Bulletin of Mathematical Biology*, vol. 51, pp. 5-37, 1989.

57. S. Wu, and U. Manber, "Fast text searching with errors," Technical Report TR91-11, Department of Computer Science, University of Arizona, Tucson, AZ, 1991.

58. R.E. Tarjan, *Data Structures and Network Algorithms*, Society for Industial and Applied Mathematics, Philadelphia, PA, 1983.

59. T.H. Cormen, C.E. Leiserson, and R.L. Rivest, *Introduction to Algorithms*, McGraw-Hill, New York, 1990.

60. R.A. Fisher, *Statistical Methods for Research workers*, Oliver and Boyd, Edinburgh, 1925.

61. M.D. Atkinson, J.-R. Sack, N. Santoro, and T. Strothotte, "Min-max heaps and generalized priority queues," *Communications of the ACM*, vol. 29, pp. 996-1000, 1986.

62. P. Wang, *An Introduction to Berkeley UNIX*, Wadsworth Publishing Company, Belmont, CA, 1988.

63. M.E. Lesk, *Lex – A Lexical Analyzer Generator*, Computing Science Technical Report No. 39, Bell Laboratories, Murray Hill, NJ, 1975.

64. S.C. Johnson, *Yacc: Yet Another Compiler Compiler*, Computing Science Technical Report No. 32, Bell Laboratories, Murray Hill, NJ, 1975.

65. W.C. Barker, D.G. George, H.-W. Mewes, and A. Tsugita, "The PIR-International protein sequence database," *Nucleic Acids Research*, vol. 20, pp. 2023-2026, 1992.

66. A. Jacobo-Molina , and E. Arnold, "HIV reverse transcriptase structure-function relationships," *Biochemistry*, vol. 30, no. 26, pp. 6351-6356, 1991.

67. J.M. Berg, "Potential metal-binding domains in nucleic acid binding proteins," *Science*, vol. 232, pp. 485-487, 1986.

68. W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling, *Numerical Recipes in C: the Art of Scientific Computing*, Cambridge University Press, Cambridge, 1988.

69. A. Bairoch, and B. Boeckmann, "The SWISS-PROT protein sequence data bank," *Nucleic Acids Research*, vol. 20, pp. 2019-2022, 1992.