Proc. Natl. Acad. Sci. USA
Vol. 89, pp. 2002–2006, March 1992
Biochemistry

Methods and algorithms for statistical analysis of protein sequences

(charge cluster/multiplet/protein periodicities/residue spacings/SAPS)

VOLKER BRENDEL, PHILIPP BUCHER†, ILLAH R. NOURBAKHSH, B. EDWIN BLAISDELL, AND SAMUEL KARLIN

Department of Mathematics, Stanford University, Stanford, CA 94305-2125

Contributed by Samuel Karlin, November 12, 1991

ABSTRACT **We describe several protein sequence statistics designed to evaluate distinctive attributes of residue content and arrangement in primary structure. Considered are global compositional biases, local clustering of different residue types (e.g., charged residues, hydrophobic residues, Ser/Thr), long runs of charged or uncharged residues, periodic patterns, counts and distribution of homooligopeptides, and unusual spacings between particular residue types. The computer program SAPS (*s*tatistical *a*nalysis of *p*rotein *s*equences) calculates all the statistics for any individual protein sequence input and is available for the UNIX environment through electronic mail on request to V.B. (volker@gnomic.stanford.edu).**

Newly derived protein sequences are commonly subjected to standard sequence analysis involving identification of similarities to other proteins in data bases, prediction of secondary structure, hydropathy plots, and mapping of potential glycosylation and phosphorylation sites and other motifs (see, e.g., refs. 1–6). Results of such primary sequence inspection may indicate structural and functional properties of the query protein and may facilitate its classification with respect to known protein families.

Here we present concepts and methods for the evaluation of a variety of protein sequence properties based on statistical criteria. Properties considered include compositional extremes, clusters and runs of charge and other amino acid types, different kinds and extents of repetitive structures, locally periodic motifs, and anomalous spacings between identical residue types. The statistics are computed for any single (or appropriately concatenated) protein sequence input to a program called SAPS (*s*tatistical *a*nalysis of *p*rotein *s*equences). The statistically significant sequence features highlighted by SAPS in the input sequence may suggest promising regions for experimental investigation. The program also finds application in the description of conserved features of families of proteins as well as in the inverse problem of deriving protein groupings based on sequence features.

Compositional extremes. Is the sequence particularly rich or poor in certain residue types relative to standard sets of proteins, grouped, for example, by species, cellular localization, function, time of expression, size, or evolutionary criteria? *Charge distribution.* Are the charged residues in the protein sequence anomalously distributed? Statistically significant clusters, runs, and periodic patterns of charge are identified. *Distribution of other amino acid types.* Does the sequence reveal significant clustering of other residue types (e.g., Ser/Thr, Cys, Pro)? Are there segments of statistically significant high hydrophobicity, and are there segments with a high propensity for forming transmembrane domains? These questions are addressed by associating appropriate scores with individual residues and locating high-scoring segments. *Repetitive structures.* SAPS identifies internal repeats (longer or more abundant than expected by chance) and high counts of multiplets (numbers of specific and nonspecific homooligopeptides). *Periodicities.* The program reports many kinds of local periodicities in the sequence that help to discern possibly regular structures. For example, repeats of period three or four often occur in amphipathic helices. Leucine every seventh residue may suggest a leucine zipper motif (7). *Spacings.* As a measure of the homogeneity of the protein, spacings between defined residue types are evaluated. For example, are the cysteines randomly spaced, or are some distances between consecutive cysteines shorter or longer than expected by chance?

The statistical tests used are described in the *Methods* section. In the *Applications* section we illustrate and interpret the output of SAPS with an analysis of the *Drosophila* neurogenic protein cut, a complex sequence that contains a variety of statistically significant features. Conservation of statistically significant sequence properties among homologous proteins is briefly discussed for the Myc family across species and cellular types.

METHODS

Compositional Analysis. Compositional extremes are evaluated relative to standard sets of proteins grouped by species, size class, cellular location, or other criteria. The standard sets are updated relative to the current release of the Swiss-Prot data base (8) and prepared in such a way that redundant entries are culled (9). The composition of the query protein is compared with the quantile table of the user-specified standard set. For example, for a set of 1154 distinct mammalian proteins of lengths between 200 and 2500 residues the median charge content per protein (combined frequency of lysine, arginine, glutamate, and aspartate) is 22.2%; 59 of these proteins have a charge content of at most 14.2% (5%-quantile point), and another 57 have a charge content exceeding 32.0% (95%-quantile point; for other examples, see ref. 10). Extremal usages that map in the tails of the reference distribution are indicated for individual amino acids and for charged and hydrophobic residues [detailed studies of compositional extremes will be presented elsewhere (S.K. and P.B., unpublished results)].

Charge Distributional Analysis. Our methods for analyzing the distribution of the charged residues in a protein sequence have been reviewed in detail previously (11). SAPS identifies charge clusters, referring to 20- to 75-residue segments with high net positive charge (positive-charge clusters), high net negative charge (negative-charge clusters), or high total charge (mixed-charge clusters) relative to the overall charge composition of the protein. The user may specify whether or not to include histidine with arginine and lysine as positively charged. For a segment of length W with a count of c charges, the program calculates the value of $t = (c - Wf)/\sqrt{Wf(1-f)}$, where f is the

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

†Present address: Institut Suisse de Recherches Expérimentales sur le Cancer, Ch. des Boveresses 155, CH-1066 Épalinges s/Lausanne, Switzerland.

fraction of charged residues in the protein. For positive- and negative-charge clusters, c is conservatively taken to be the net count of positive or negative charges in the window and f is the fraction of positive or negative charges in the protein, respectively. Following a binomial model of successes and failures in W trials, the count c is deemed significant for large t. To accommodate the problem of multiple comparisons (tests), the criterion for t is adjusted according to the length of the protein. Conservatively, only clusters with t values exceeding 5 for proteins longer than 1500 residues, 4.5 for proteins of lengths between 750 and 1500 residues, and 4 otherwise are printed. Algorithmic considerations are discussed in ref. 11. Clusters are also evaluated by scoring schemes (e.g., scores associated with positive-charge clusters assign +2 to arginine and lysine, −2 to glutamate and aspartate, and −1 to all other residues; see below for the evaluation of high-scoring segments).

The program further identifies significantly long runs of charged residues (i.e., contiguous occurrences of charge residues, typically six or more residues long allowing for a few interruptions of noncharged residues counted as errors), runs of uncharged residues, and periodic patterns of charge (e.g., charged residues every second or every third residue). The significance of charge runs within a protein is estimated by reference to a random sampling model. For a sequence having length N and a letter occurring with frequency f, the probability of observing a run of this letter of length exceeding $L = \ln N/(-\ln f) + z$ is asymptotically at most $1 - \exp\{-(1-f)f^z\}$ (12). Setting this probability equal to 0.01 we obtain z and L corresponding to the length of runs significant at the 1% level. Formulas for estimating the significance of runs with errors and of periodic patterns are given in ref. 11.

Distribution of Other Amino Acid Types. This section of SAPS reveals local clustering of amino acid types other than charge. Local clustering is evaluated by a scoring method (13). SAPS routinely scores for hydrophobic segments and for transmembrane domains. Scores associated with hydrophobicity were chosen in rough proportion to standard hydrophobicity indices (e.g., those given in ref. 3): −8 for K, E, D, R; −4 for S, T, N, Q; −2 for P, H; +1 for A, G, Y, C, W; and +2 for L, V, I, F, M. The protein is scanned and scores are accumulated according to the sequence. Segments of high cumulative score correspond to hydrophobic regions. Statistical significance is determined as described elsewhere (13).

Scores for the detection of transmembrane domains were obtained as follows: All sequences in Swiss-Prot Release 17 with annotated transmembrane regions were compiled and split into (putative) transmembrane and complementary parts. The residue compositions in these aggregate sequences were derived for a reduced 11-letter alphabet (see below; frequencies q_i and p_i, respectively). Then integer scores were chosen proportional to $\ln(q_i/p_i)$ to obtain −8 for E, D; −7 for K, R; −4 for P, N, Q; −3 for H; −1 for S, T; +1 for A, G, Y, M, C, W; and +3 for L, V, I, F. These scores are best suited for the purpose of associating high-scoring segments with target domains of standard transmembrane composition (cf. refs. 10 and 13) different from hydrophobic cores of globular domains.

SAPS also identifies clusters of user-specified residue types. Statistical significance is established in two ways, as for charge clusters, following a binomial model of occurrences of successes and failures (11) and high-scoring segment statistics (13), respectively. Different scoring schemes are applied to screen for various allowances of errors. For example, to test for cysteine clusters, a score of +2 is associated with C, and scores of alternatively −10, −7, −4, or −1 are given to all other residues; the more-negative scores give higher penalties for noncysteine residues and thus are suitable for detecting very high local concentrations of C (essentially runs), while the scoring schemes involving less negative scores for noncysteine residues screen for more diffuse cysteine clusters. Thresholds for statistically significant positive aggregate scores are determined as described in ref. 13.

Repetitive Structures. All internal repeats exceeding some minimal length (allowing for error extensions) are printed. Repeats are displayed in blocks such that repeats of copy number higher than two are put together. The repeat algorithm follows essentially the method of Leung *et al.* (14) for multiple sequence comparisons and is implemented for both the full 20-letter amino acid alphabet and for a reduced 11-letter alphabet that groups the hydrophobics (L, V, I, F); the charged residues (R, K and E, D); the small amino acids (A, G); as well as S, T (hydroxyl group); N, Q (amide group); and Y, W (aromatic); with C, H, M, and P left separate.

Repetitiveness is scrutinized further by means of *multiplet* counts. Here multiplet stands for any homooligopeptide (e.g., A_2, Q_7). Extremes in these counts are determined as follows. Let f_i be the frequency of residue type i in the sequence. In a random sequence, a given site is the first residue of a multiplet of residue type i with probability $p_i = \sum_{k=2}^{\infty} f_i^k(1-f_i)^2 = (1-f_i)f_i^2$. Consequently, the probability of observing a multiplet of any amino acid is given by $p = \sum_{i=1}^{20}(1-f_i)f_i^2$. If N denotes the length of the sequence, then we take the aggregate multiplet count (combined number of all homooligomers) to be significant if the count exceeds $Np + 3\sqrt{Np(1-p)}$ (15). A significant multiplet count for amino acid i is required to exceed $Np_i + 4\sqrt{Np_i(1-p_i)}$; here 4 SDs above the mean is chosen as a conservative significant threshold to accommodate the problem of testing all 20 amino acids simultaneously. SAPS similarly determines exceptional counts of specific *altplets*, referring to runs alternating in amino acids i and j (e.g., KPKPK). In this case p_i is replaced by $p_{ij} = f_if_j(2 - f_i - f_j)$. Multiplets and altplets are also evaluated in the charge alphabet.

Periodicity Analysis. The query protein is screened for periodic repeats. A periodic pattern is represented by the starting position, a string characterizing the repeat unit, and the number of repetitions. The string characterizing the repeat unit identifies amino acids in the one-letter code (or +, −, and ∗ for positive, negative, and either charge, respectively); dots denote variable positions. The subroutine generates a complete nonredundant list of periodic patterns. All periodic patterns of period at most 10 and repeat counts of at least 5 are printed; for periods 1 (runs) and periods 2, 3, 4, and 7 (potentially relevant to β-sheet or α-helical structures) the required minimum count is lowered to 4. For a random sequence of length N the probability of observing a run of repeat count r and given period of a specified letter occurring with frequency f is approximated by $1 - (1-fx)/[(r+1-rx)(1-f)x^{n+1}]$, where x solves $(1-f)x(1 + fx + \cdots + f^{r-1}x^{r-1}) = 1$ (16). With $f = 0.05$ and r replaced by $r-1$ in these formulas one obtains a lower bound for the probability of observing an r repeat of any (not predetermined) amino acid, which indicates a minimum value of r that might convey statistical significance. In the SAPS output, periodic patterns of amino acids or of charged residues are marked if their lengths are sufficient to give probabilities of roughly 0.01 in the above sense.

Spacing Analysis. A set of n marked residues in a sequence of length N (e.g., all cysteines or all basic residues) induces $n+1$ spacings from the N-terminus of the protein to the first marker, from the first marker to the second, and so on (10). SAPS determines the maximal and minimal spacings for all amino acids and charges and displays statistically significant spacings. Unusually long maximal spacings and unusually short minimal spacings suggest an inhomogeneous distribution of the respective amino acid or charge type. In the same way, unusually short maximal spacings and unusually long minimal spacings indicate excessive evenness in the positioning of the residues.

Table 1. Compositional analysis of the *Drosophila* cut protein

a.a.	Count (frequency)	a.a.	Count (frequency)	a.a.	Count (frequency)
A+	294 (13.5%)	C	8 (0.4%)	D	114 (5.2%)
E	150 (6.9%)	F	39 (1.8%)	G	127 (5.8%)
H	85 (3.9%)	I	54 (2.5%)	K	88 (4.0%)
L	182 (8.4%)	M	58 (2.7%)	N	144 (6.6%)
P	129 (5.9%)	Q	203 (9.3%)	R	83 (3.8%)
S	215 (9.9%)	T	104 (4.8%)	V	66 (3.0%)
W	10 (0.5%)	Y−	22 (1.0%)		

a.a., Amino acid. For the sequence of the cut protein, see Swiss-Prot entry HMCU$DROME or ref. 18. Other counts (frequencies) are as follows: positive residues, 171 (7.9%); negative residues, 264 (12.1%); total charges, 435 (20.0%); net charge, −93 (−4.3%); hydrophobic content, 399 (18.3%). Frequencies are evaluated with respect to a reference set of 242 *Drosophila* proteins. +, High use of alanine (exceeding the 95%-quantile point of the reference set); −, low use of tyrosine (below the 5%-quantile point of the reference set).

For statistical evaluation, spacings between amino acids are interpreted as runs of failures (designated F) between successes (designated S). Then, the probability that the kth longest spacing is greater than or equal to s is given by the probability that a sequence of n successes and $(N - n)$ failures contains at least k runs of failures greater or equal to s. The latter probabilities will be given elsewhere (M. Morris, G. Schachtel, and S.K., unpublished results). A maximal spacing ($k = 1$) is displayed if it is matched or exceeded with probability <0.01 or >0.99, in which case also the second maximal spacing is displayed if its probability is <0.05 or >0.95. The minimal spacing is evaluated by an approximate formula. Let $x = (i - j)/N$, where i and j are the coordinates of the minimally spaced markers. Then $p = [1 - x(n + 1)]^n$ approximates the probability that the minimum for a random sequence model is at least x (17). Minimal spacings for amino acids comprising <4% of the protein are displayed if p is <0.01 or >0.99.

APPLICATIONS

To illustrate the output of SAPS we concentrate on a detailed discussion of the 2175-amino acid *Drosophila* cut sequence (18). cut is a homeodomain-containing polypeptide that derives from a genomic region of more than 70 kilobases at the *cut* locus. The expression of *cut* appears to direct development of external sensory organs (18). We chose cut as an example because of its extraordinary richness and variety in significant sequence features. It should be noted, however, that most proteins would yield much sparser annotations; for example, globins, actins, and many enzymes are devoid of any statistically significant sequence features as defined in *Methods*.

Compositional Analysis. The cut residue composition is evaluated with respect to a collection of 242 *Drosophila melanogaster* proteins. cut is rich in alanine (exceeding the 95%-quantile of the reference set) and poor in tyrosine (Table 1). The alanines form several small clusters (evaluated by scoring statistics as discussed below) located throughout the sequence, including 10 runs each longer than five residues (see Table 3). The tyrosines occur predominantly in the second two-thirds of the protein as there is only one tyrosine in the first 835 residues (revealed by the spacing analysis, Table 5, discussed below).

Charge Distributional Analysis. cut displays two distinct acidic-charge clusters followed by two mixed-charge clusters (Table 2). Proteins with multiple charge clusters are rare (<3% of mammalian proteins, ref. 19). It is of note that the cut homeodomain (residues 1744–1804, characterized as a divergent member of the homeodomain set, ref. 18) is not significantly charged, in contrast to most homeodomains, which generally qualify as significant mixed-charge clusters (20). The SAPS output further identifies significantly long runs and periodic patterns of charge. cut contains a plethora of significant charge runs, most of which coincide with or are contained within the charge clusters. Five significantly long runs of uncharged residues occur in locations dispersed throughout the sequence (Table 2). Charge analysis by scores gives similar results and is not shown.

The first mixed charge cluster in Table 2 contains the remarkable iteration of RE five times with one interrupting Q, a statistically highly significant alternating run. Long alternating charge runs also occur in the nuclear ribonucleoprotein U1-70K and in the mouse major histocompatibility complex class III 42-kDa polypeptide; see ref. 21.

The two longest acidic runs in cut (length 22 with one error beginning at residue 271, and length 9 with no error beginning at residue 574; Table 2) qualify as hypercharge runs (Table 2). Hypercharge runs refer to extremely long charge runs judged both by protein internal standards as described in *Methods* and by comparison with an average protein residue composition (e.g., hyperacidic runs are required to exceed the size that would occur with probability 10^{-5} in a random sequence of the same length with 11.5% acidic residues, ref. 15). Such runs are prominent in nuclear autoantigens (21).

Distribution of Other Amino Acid Types. To illustrate the use of scoring schemes to evaluate clustering of particular amino acids we have analyzed the alanine distribution in cut. The SAPS program allows the user to optionally analyze the distribution of any other amino acid (or of combinations, e.g., S, T) in the same way. Given scores 2 for alanine and −1 for all other residues, the 5% and 1% significance thresholds for segment scores are calculated to be 13 and 15, respectively (see *Methods* and ref. 13); i.e., segments of cumulative score exceeding these thresholds signify nonrandom clustering of the alanines in cut. There are six such segments (of lengths at least 10 residues, a restriction generally set in SAPS to avoid duplicate printing of error-free runs), starting at residues 74 [length (L) 19, score (S) 20], 228 (L, 16; S, 20), 616 (L, 15; S,

Table 2. Charge distributional analysis of the cut protein

Clusters and runs				Uncharged runs		
Type	Location	Sequence	t value*	Location	Length†	Composition‡
Negative	271–293	**DEEELDDEEEDEEEDEDEDDEEE**	12.26	120–179	59 (1)	L, Q, S, A, N
	547–582	**EDDEEDEDQAMLVDSEEAEDKPEDSHHDDDEDEDED**	9.51	664–782	117 (2)	H, Q, S, N
Mixed	1063–1085	**RERERE**QR**RERE**QQQRLRHDDQDK	5.94	1231–1284	54 (0)	P, A, G, S
	2033–2111	**RDDTDSNKPTDGGNDSDHEHAQLEIDQRFMEPEVHIKQEE-DDDEE**QSGSVNLDNE**D**NATSEQKLKVINEEKLRMVRVRR	5.40	1459–1681	122 (1)	Q, A

Charged residues are shown in boldface type. Hyperacidic runs are underlined. Residues 418–422 (KRKKK) of cut comprise a significant positive charge run not shown here.

*The 1%-significance threshold is 5.0 (see text).

†Numbers in parentheses are numbers of intervening charged residues.

‡Residues each comprising >10% of the run.

24), 794 (L, 14; S, 13), 1876 (L, 13; S, 23), and 2004 (L, 11; S, 19).

Repetitive Structures. cut contains a 60-amino acid triple repeat (residues 886–945, 1339–1398, 1617–1676) described previously (18), the core identity blocks of which are displayed in the SAPS output (not shown). Assessment of multiplet counts affords a different notion of the repetitiveness of a sequence. In cut there is a striking overrepresentation of multiplets (247 observed, with the significance threshold at 178; Table 3). Individual amino acids of significant multiplet count are glycine, glutamine, and histidine (Table 3). About 12% of *Drosophila* proteins have significant multiplet counts, as compared to only 2% of mammalian and yeast proteins and <0.5% of *Escherichia coli* proteins (data not shown). No general role for such an abundance of reiterated residues (mostly doublets and triplets, separated by small spacers of <10 residues) is known. It may be functional in increasing the size of the protein and providing space filling between different domains. cut, like many other *Drosophila* developmental proteins, also contains several long isolated multiplets (Table 3), corresponding predominantly to DNA tandem repeats (18).

Periodicity Analysis. The periodicity output of SAPS is illustrated in Table 4 (to save space, printing of periodicities of period 1, i.e. runs, has been suppressed). For periodic elements involving several conserved positions a consensus is derived and displayed with the number of errors in each position indicated.

Spacing Analysis. Several spacing anomalies are evident in cut (Table 5). For example, the largest spacing between tyrosines is 803 and refers to the tyrosine-free region between residues 33 and 836. In a random sequence such a long gap would be expected only with probability 0.0009. On the other hand, the second largest spacing between tyrosines is rather small compared to what is expected by chance ($P = 0.98$), further underlining the nonhomogeneous distribution of the tyrosines. The excessively long spacings indicated for the charges correspond to the previously identified uncharged runs.

DISCUSSION

We have presented the SAPS computer program, which evaluates a variety of statistical properties of protein sequences. These properties pertain to significant compositional biases; clusters, runs, and periodic patterns of charged residues; several forms of repetitive elements; high-scoring segments for hydrophobicity and propensity for transmembrane location; general amino acid periodicities; and unusual spacings between different residue types. Applications of SAPS include computer-guided experimental designs for the investigation of protein structure and function, characterization of protein families, and protein classification.

Advances in recombinant DNA and sequencing technology have made elucidation of the primary sequence of a protein a relatively easy and early step in the description of the protein's structure and function. Once the sequence is obtained, experimental investigation will focus on the characterization of functional domains and their interactions. The SAPS output is designed to help identify likely regions of structural or functional importance for further experimental study. For exam-

Table 3. Multiplet analysis of the *Drosophila* cut protein

Amino acid multiplets	
Total number, 247 (critical number, 178)	
Significant individual multiplet counts	
Amino acid	Observed (critical) number
G	19 (18)
Q	38 (34)
H	12 (11)
Multiplets exceeding length 5 (letter/length/position)	
A/5/75, A/9/194, Q/7/204, A/9/235, N/7/404, N/6/423, N/5/506, A/11/618, Q/5/1150, A/5/1185, A/7/1530, A/5/1876, A/7/1882, A/5/1967, A/8/2007, P/6/2129, S/6/2148	
Charge multiplets	
Total number, 56 (critical number, 60)	
19 +plets (7.9%), 37 −plets (12.1%)	
49 altplets, (critical number, 62)	

Critical numbers for multiplet counts were established as described in the text. Also shown are all multiplets exceeding length 5. Charge altplets refer to alternating reiterations of positively and negatively charged residues.

Table 4. Periodic elements in the *Drosophila* cut sequence

Period	Element	Position	Copies (errors)*
2	A.	74–93	9 (1)
5	E. . . .	273–297	5
2	Q.	321–328	4
3	–0.	328–342	5
9	N.	388–432	5
2	N.	402–411	5
4	–. . .	550–585	8 (1)
4	D. . .	566–585	5
2	A.	616–631	8
4	Q. . .	669–688	5
7	Q.	670–704	5
7	QR.R. . .	1062–1089	4 (0, 1, 1)
3	P. .	1235–1246	4
2	G.	1243–1250	4
5	T. . . .	1301–1325	5
6	Q. . Q. .	1464–1517	8 (1, 3)
3	Q. .	1482–1523	12 (2)
2	Q.	1482–1515	13 (4)
4	Q. . .	1497–1532	8 (1)
2	A.	1965–1972	4

Periodic elements are identified by consensus (one-letter code and charge alphabet) and variable (".") positions.
*Numbers in parentheses are numbers of mismatch errors (if any) in each of the consensus positions.

Table 5. Significant large spacings in the *Drosophila* cut sequence

	Position (x)	Spacing*	P†
	Amino acid		
Y	33	803	0.0009
Y	1429	179	0.9787
R	42	271	0.0011
R	643	190	0.0003
F	183	453	0.0040
F	1401	239	0.0415
G	429	226	0.0001
G	1438	110	0.0066
D	1404	212	0.0007
D	118	153	0.0001
N	1406	166	0.0010
N	1813	88	0.0318
	Type of charge		
+	653	130	0.0029
+	1458	125	<0.0001
−	1461	121	<0.0001
−	1958	57	0.0084
+ or −	1461	121	<0.0001
+ or −	1230	55	<0.0001

*Distance between the residue at position x and the next occurrence of the same residue in the sequence. Shown are the maximal and second maximal spacings.
†Calculated as described in the text.

Table 6. Hyperacidic charge runs and other features of Myc proteins

Protein	Length*	Sequence with flanks	Position†	bHLH position‡	Extremal a.a. usage§	Significant multiplets
Human c-Myc	439	TSS **DSEEEQEDEEEID** VVS	251	355	S+, G−	Yes
Chicken c-Myc	416	TSS **DSEEEQEEDEEID** VVT	228	332	G−	Yes
Xenopus c-Myc1	419	SSS **ESEEEPEDEDEDCDEEEEID** VVT	226	336	S++, T−, G−−	No
Trout c-Myc	(414)	SGS **DSEDDDEEEDDDEDEEEID** VVT	220	325	S+, D+, M−, G−−	No
Human N-Myc	464	TLS **DSDDEDDEEEDEEEEID** VVT	262	382		Yes
Chicken N-Myc	441	TLS **DSDDEDEEEEDDEEEID** VVT	241	359		Yes
Mouse L-Myc	368	APK **EKEEEEEEEEEEE** IVS	245	286	L−	No

One mammalian Myc protein per subtype is shown. Charged residues are shown in boldface type. The hyperacidic charge runs of four other sequenced mammalian Myc proteins are identical in the charge alphabet relative to subtype; woodchuck N-Myc2 contains the run **DEVDEEEDEEEEID**, and human L-Myc contains **EKEDEEDEE**. All sequences were taken from Swiss-Prot (8).

*The trout c-Myc sequence is incomplete.

†Position of the first amino acid of the displayed hypercharge run.

‡Position of the first conserved basic residue of the helix–loop–helix domain according to the alignment of Benezra *et al.* (22).

§Extremal amino acid (a.a.) usage is based on quantile tables for 798 distinct human proteins: ++, 99–100% quantile range; +, 95–99% quantile range; −, 1–5% quantile range; −−, 0–1% quantile range.

ple, both the *Drosophila* cut protein and all the Myc proteins contain exceptionally long runs of acidic residues (Tables 2 and 6). Site-directed mutagenesis or truncation and fusion experiments may reveal a role for these outstanding sequence features.

As an example of characterization of a protein family by unique statistical sequence features we discuss the Myc protein family. The Myc proteins are nuclear phosphoproteins of the helix–loop–helix protein class (22). Myc proteins appear to function in cell growth control and are implicated in neoplasia, but their exact role is as yet unclear. All of the Myc proteins carry a hyperacidic run (Table 6), although this run is neither conserved in amino acid identity nor in location (varying between 41 and 120 residues amino proximal to the basic region that precedes the helix–loop–helix motif). This feature is unique to the Myc family among helix–loop–helix proteins and other transcription factors and oncogenic proteins (23). It is conceivable that anionic repulsion due to the acidic region functions in preventing Myc dimerization or helps orientation of Myc monomers in other oligomeric complexes or with respect to DNA binding sites. The c-Myc proteins are universally rich in serine and low in glycine (Table 6). The majority of Myc proteins have significantly high aggregate multiplet counts, a property not shared by most other helix–loop–helix proteins (E12, MyoD, achaete-scute, daughterless, and others) and generally rare (occurring in <2% of mammalian proteins; data not shown).

Systematic studies of the properties evaluated by SAPS may help in efforts to classify proteins and to develop protein systematics based on sequence similarities. For example, significant charge configurations have very different prevalence for different classes of oncogene products (23). Extensive studies of compositional biases in proteins of different types will be presented elsewhere (S.K. and P.B., unpublished results).

SAPS is written in the C programming language for the UNIX environment and is available on request from V.B. (electronic mail address: volker@gnomic.stanford.edu).

We wish to thank Drs. Ron Sapolsky and Michael Zuker for critical comments on the manuscript. This work was supported by National Institutes of Health Grants HG00335-04 and GM10452-28 to S.K.

1. Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 2444–2448.
2. Biou, V., Gibrat, J. F., Levin, J. M., Robson, B. & Garnier, J. (1988) *Protein Eng.* **2,** 185–191.
3. Kyte, J. & Doolittle, R. F. (1982) *J. Mol. Biol.* **157,** 105–132.
4. Bause, E. (1983) *Biochem. J.* **209,** 331–336.
5. Krebs, E. G. & Beavo, J. A. (1979) *Annu. Rev. Biochem.* **48,** 923–959.
6. Bairoch, A. (1991) *Nucleic Acids Res.* **19,** 2241–2245.
7. Landschulz, W. H., Johnson, P. F. & McKnight, S. L. (1989) *Science* **243,** 1681–1688.
8. Bairoch, A. & Boeckmann, B. (1991) *Nucleic Acids Res.* **19,** 2247–2249.
9. Brendel, V. (1992) in *Advances in Mathematics and Computers in Medicine*, ed. Witten, M. (Pergamon, New York), Vol. 5, in press.
10. Karlin, S., Bucher, P., Brendel, V. & Altschul, S. F. (1991) *Annu. Rev. Biophys. Biophys. Chem.* **20,** 175–203.
11. Karlin, S., Blaisdell, B. E. & Brendel, V. (1990) *Methods Enzymol.* **183,** 388–402.
12. Karlin, S., Ost, F. & Blaisdell, B. E. (1989) in *Mathematical Methods for DNA Sequences*, ed. Waterman, M. (CRC, Boca Raton, FL), pp. 133–157.
13. Karlin, S. & Altschul, S. F. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 2264–2268.
14. Leung, M.-Y., Blaisdell, B. E., Burge, C. & Karlin, S. (1991) *J. Mol. Biol.* **211,** 1367–1378.
15. Karlin, S., Brendel, V. & Bucher, P. (1992) *Mol. Biol. Evol.* **9,** 152–167.
16. Feller, W. (1968) *An Introduction to Probability Theory and Its Applications* (Wiley, New York), 3rd Ed., Vol. 1, p. 325.
17. Karlin, S., Burge, C. & Campbell, A. M. (1992) *Nucleic Acids Res.*, in press.
18. Blochlinger, K., Bodmer, R., Jack, J., Jan, L. Y. & Jan, Y. N. (1988) *Nature (London)* **333,** 629–635.
19. Karlin, S. (1990) in *Structure and Methods*, eds. Sarma, R. H. & Sarma, M. H. (Adenine, Albany, NY), Vol. 2, pp. 171–180.
20. Brendel, V. & Karlin, S. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 5698–5702.
21. Brendel, V., Dohlman, J., Blaisdell, E. B. & Karlin, S. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 1536–1540.
22. Benezra, R., Davis, R. L., Lockshon, D., Turner, D. L. & Weintraub, H. (1990) *Cell* **61,** 49–59.
23. Karlin, S. & Brendel, V. (1990) *Oncogene* **5,** 85–95.