Numerical classification of coding sequences

David W. Collins, Chia-Chang Liu and Thomas H. Jukes

Space Sciences Laboratory and Department of Integrative Biology, University of California, Berkeley, CA 94720, USA

Received September 10, 1991; Revised and Accepted February 14, 1992

ABSTRACT

DNA sequences coding for protein may be represented by counts of nucleotides or codons. A complete reading frame may be abbreviated by its base count, e.g. $A_{76}C_{158}G_{121}T_{74}$, or with the corresponding codon table, e.g. $(AAA)_0(AAC)_1(AAG)_9...(TTT)_0$. We propose that these numerical designations be used to augment current methods of sequence annotation. Because base counts and codon tables do not require revision as knowledge of function evolves, they are well-suited to act as cross-references, for example to identify redundant GenBank entries. These descriptors may be compared, in place of DNA sequences, to extract homologous genes from large databases. This approach permits rapid searching with good selectivity.

INTRODUCTION

The annotation of DNA sequences in databases needs consideration and, we submit, supplementation, in the light of their present cataloging in GenBank and elsewhere. This viewpoint is emphasized by the existence of identical sequences listed under different names. We have found many examples of this, so that the actual number of different genes is considerably smaller than the number of described reading frames. Current indexing methods do not permit these duplicates to be detected easily. A related problem is the inconsistent description of related (*i.e.* orthologous) sequences from different organisms. Accordingly, it is often necessary to perform a computationally intensive search of DNA sequences in order to selectively extract homologous sequences.

These problems are likely to persist as the databases continue to expand. The GenBank (1) nucleic acid sequence database (release 69.0, 1991) describes 55,631 loci representing 71,947,426 bases and is 9.0% larger (in bases) than the previous release. The total size of the database, including indices, is nearly 300 megabytes and it is growing at a nearly exponential rate (2). As efforts to sequence the human and other genomes proceed, this growth is expected to continue. Managing this information and making it easily accessible has been identified as one of the major challenges of the human genome project. Now that nucleotide sequences of thousands of genes are available, it is timely to consider additional methods of classification.

In the present communication, we evaluate two simple methods for specifically characterizing protein-coding sequences. The first is the base count. All nucleotide sequences can be described by

four independent variables: the numbers of each of the four bases A, C, G and T(U). As an example, the human zeta-globin gene sequence (3), excluding introns and including initiation and stop codons, contains 429 nucleotides distributed as $A_{76}C_{158}G_{121}T_{74}$. Like a molecular formula, this designation lists the number of each component, but contains no information about their arrangement. A second, more informative, description is the codon table, a list of the frequencies of each of the 64 codons. A codon table provides the amino acid composition of the corresponding protein and the pattern of synonymous codon usage of the gene.

RESULTS AND DISCUSSION

Limitations of Functional Annotation

Gene and DNA sequence nomenclature is primarily based on knowledge of function. For example, the Guidelines for Human Gene Nomenclature (4) have recommended that '*the name of a gene...should convey information about the character or function of the gene. The name may indicate, for example, a morphological or disease character when the gene function is not known, or a biochemical, antigenic, or molecular property. Ultimately a gene name should evolve to indicate the exact gene function.*'. Human loci are typically represented by a few alphanumeric characters chosen to abbreviate function, *e.g.* *G6PD*, *HPRT*, and *HBB* for glucose-6-phosphate-dehydrogenase, hypoxanthine-phosphoribosyl transferase, and hemoglobin beta respectively (4). Sequence files in GenBank are usually described according to functional criteria. Annotated sequences are described with a LOCUS name, a few characters chosen to abbreviate the source organism and function, a DEFINITION, a further brief description of function, and KEYWORDS. Each file is identified with an ACCESSION number, a code that carries no information about the associated data (2).

There are several limitations associated with using function-based descriptions by themselves to identify gene sequences. First, a sequence may have more than one function. For example, the metabolic enzyme lactate dehydrogenase B₄ and the structural protein epsilon-crystallin of lens are the same protein in duck, coded by a single gene (5). In these cases it may not be obvious which function should be given priority in classification. Second, a gene's function may be poorly understood or even unknown even after its nucleotide sequence has been determined. Consequently, descriptions based on

function are subject to change as knowledge is acquired, necessitating frequent revision of the corresponding nomenclature. The human p53 tumor suppressor gene has at different times been classified as a tumor antigen, oncoprotein, and tumor suppressor (6). The mouse gene *int-1* was renamed *wnt* to avoid confusion with two unrelated genes, *int-2* and *int-3* (7). Recently, several hundred cDNAs from human brain were partially sequenced, including many of unknown function (8). Efforts to sequence the human and other genomes will produce many more sequences whose function cannot be immediately described. Third, functional terms and abbreviations are difficult to standardize and apply consistently. For example, terms such as 'protease' and 'proteinase' are often used interchangeably. Zeta-globin is abbreviated HBA1, XIHB, AZGLO, and ZGL in GenBank LOCUS names for this sequence from human, chimp, and horse. Fourth, a large number of sequences may be found to have the same or similar function, and may not be easily distinguished on this basis. It has been predicted that the human genome may code for a thousand kinases and a thousand phosphatases (9). As the databases continue to grow, the utility of terms such as 'kinase' for database queries will be diminished. Finally, descriptions based on function convey but little information about the presence or absence of evolutionary relationships, which may be of great interest.

Database Redundancy and Variations in Annotation

There are many possible ways to describe the same sequence in functional terms and these descriptions, for the reasons mentioned above, often do not converge. Not surprisingly, databases such as GenBank contain many redundant entries, *i.e.* the same (or nearly identical) gene sequence may be present in more than one file with different annotation. Wada *et al.* (10) extracted codon tables for 1,490 human sequences from release 62.0 of GenBank. Of these, 365 duplicate another codon table with 64, 62, 60 or 58 matches, suggesting sequences that differ by only 0, 1, 2, or 3 nucleotide substitutions, respectively. This indicates that about 25% of these entries are redundant in the

sense that the same coding sequence is associated with two or more LOCUS names.

For example, we have found one case where an identical 963 bp human coding sequence (base count A₂₈₁C₂₁₀G₂₇₂T₂₀₀) has been deposited independently in GenBank six times (ACCESSION M18366, D00172, J03745, M21731, M19384, X12454). It is variously defined as endonexin II (LOCUS HUMENN), lipocortin-V (HUMLC5), blood coagulation inhibitor (HUMBCI), placental anticoagulant protein PAP (HUMATC), placental anticoagulant PP4 (HUMPAP4), and vascular anticoagulant (HUMVAC) (11, 12, 13, 14, 15, 16). Endonexin II belongs to a family of Ca²⁺- and phospholipid-binding proteins additionally referred to as synexins, annexins, chromobindins, calmedins, calpactins, calelectrins, and proteins I-III (14, 17). It has recently been shown to form a Ca²⁺ channel in lipid bilayers (17); inhibition of blood coagulation inhibition may not be a physiological function.

Duplicate entries such as these are difficult to detect on the basis of annotation. GenBank provides indices based on accession number, keyword phrase, author name, journal citation, and gene symbol (2). These may be insufficient for locating related sequences, because, as the previous example demonstrates, even identical nucleotide sequences may be differently annotated. For example, to retrieve all six files containing the endonexin II sequence with a search of the keyword index, the minimum set of keywords needed is 'anticoagulant', 'coagulation', 'endonexin', and 'lipocortin'. Even if this combination could be anticipated in advance, such a search has poor selectivity; many additional unwanted records are retrieved. Alternatively, one may scan the DNA sequences themselves, rather than the associated annotation, but this approach is computationally expensive. Nucleotide sequences typically contain far more information than is needed to uniquely select a gene from a database and have variable lengths, necessitating alignment before comparison.

A related problem is the extraction of homologous sequences from different organisms. Homologous genes are even less likely to be described consistently than are multiple submissions of the same gene. Table 1 lists some examples of closely related human and rodent gene sequences with different GenBank definitions.

Table 1. Homologous human and rodent genes may have different descriptions. GenBank LOCUS names and definitions of six pairs of homologous human and rodent sequences are listed below.

ACCESSION	LOCUS	DEFINITION
1) J03189	HUMSECT	Human proteolytic serine esterase-like protein
M12302	MUSCCPA	Mouse C11 mRNA encoding T-cell specific protein CCPI
2) X15187	HUMTRA1	Human glycoprotein 96 (tra1)
J03297	MUSERPX	Mouse ERp99 mRNA encoding an endoplasmic reticulum transmembrane protein
3) J03779	HUMCALLA	Human common acute lymphoblastic leukemia antigen (CALLA)
M15944	RATENKA	Rat enkephalinase (neutral endopeptidase)
4) X07820	HUMSTROM2	Human mRNA formetalloproteinase stromolysin 2
X02601	RATPTR1	Rat pTR1 mRNA encoding a 53 Kd protein
5) M14043	HUMLIC	Human lipocortin II
M14044	MUSCALP	Mouse calpactin I heavy chain (p36)
6) M18366	HUMATC	Human placental anticoagulant protein (PAP)
M21730	RATLC5	Rat lipocortin-V

Base Counts and Codon Tables

As a supplement to functional annotation, we turned to one of the simplest characterizations of a coding sequence, its nucleotide content, and found that this characterization gives a concise and surprisingly distinctive identification of a gene's coding sequence. Identical sequences necessarily have the same base counts, but

Table 2. Sequences having 0, 1, 2, 3, or 4 matches with the base count of the human zeta-globin gene. Human zeta-globin base count, A₇₆C₁₅₈G₁₂₁T₇₄, was compared to those of 9,037 complete coding sequences from GenBank 62.0. GenBank LOCUS names HUMXIHB and HUMHBA1 both represent the human zeta-globin sequence, HUMHBA4 #1 and HUMHBA4 #2 the (identical) human alpha-1- and alpha-2-globin genes, and ECORPSFRI #2 an unidentified reading frame of *E. coli*. ECORPSFRI #2, A₇₆C₇₈G₈₇T₇₄, has matching counts of A and T, HUMHBA4 #1 and HUMHBA4 #2 (both A₇₇C₁₅₇G₁₂₁T₇₄), have the same counts of G and T.

Matches	Sequences	GenBank LOCUS
0/4	8,921	
1/4	111	
2/4	3	HUMHBA4 #1, HUMHBA4 #2, ECORPSFRI #2
3/4	0	
4/4	2	HUMHBA1, HUMXIHB

the opposite is not true; *i.e.* sequences having the same base counts are not necessarily identical.

In order to estimate the likelihood of a coding sequence having the same base count as unrelated sequences in the databases, we compared the base count of the human zeta-globin gene, A₇₆C₁₅₈G₁₂₁T₇₄, to those of 9,037 other complete reading frames. Base counts were derived from codon usage tables extracted from release 62.0 of GenBank by Wada *et al.* (3). Genes corresponded to the primate (PRI), rodent (ROD), other mammal (MAM), other vertebrate (VRT), invertebrate (INV), plant, fungi, algae (PLN), organelle (ORG), and bacteria (BCT) divisions of GenBank. Sequences from the viral (VRL), phage (PHG) and unannotated (UNA) divisions of GenBank were not included. All comparisons were done using microcomputers. For each comparison, the number of matches, *i.e.* either A = 76, C = 158, G = 121, or T = 74, with the human zeta-globin base count was tabulated.

Table 2 summarizes the results. Remarkably, only 111 (1.23%) of the 9,037 comparisons produced the same count of any one base. Only three (0.033%) have the same counts of 2 bases. Two of these three are for homologous human alpha-1 and alpha-2 globins, the third is for an unidentified reading frame of *E. coli*. None have identical counts of 3 of four bases. When the numbers of all four bases are the same, the sequences are identical as well.

Based on Table 2, the chance of a pair of unrelated genes having the same count of any of the four bases can be roughly

estimated as 1 in 10² and for two of the four, 1 in 10⁴. The chance of a given pair of unrelated genes having the same count of all four bases is about 1 in 10⁸. The base count is therefore sufficient to winnow out a coding sequence from millions of others.

Sequences that are nearly identical, *i.e.* those differing by point mutations or sequencing errors, necessarily have similar base counts and may be recognized on this basis. Accordingly, base counts may be used to selectively retrieve closely related sequences from databases. We compared a base count for the human p53 tumor suppressor gene (18), A₂₇₆C₃₆₄G₃₀₈T₂₃₄, to 9,037 others. Each comparison was assigned a score equal to the Euclidean distance between base counts: [(A-276)² + (C-364)² + (G-308)² + (T-234)²]^{1/2}. The 10 lowest scores are listed in Table 3 with the corresponding base counts. The five lowest scores, all associated with human p53 sequences, are easily distinguished from the bulk of the comparisons.

Codon tables may be used to identify sequences in much the same way. Codon tables for the zeta-globin genes of human, chimp and horse are given in Table 4. The three sequences are homologous, relatively short (143 codons), and have a pattern of synonymous codon choice that is biased towards G and C in codon third positions. Nevertheless, the three genes can be easily differentiated on the basis of codon usage. The human and chimp tables differ in usage of 10 of 64 codons, the human and horse differ in usage of 30 of 64 codons.

Table 3. Selective retrieval of similar nucleotide sequences using base counts. A base count for human p53 tumor suppressor gene, A₂₇₆C₃₆₄G₃₀₈T₂₃₄, was compared to 9,037 others from GenBank 62.0 using the formula Score = [(A-276)² + (C-364)² + (G-308)² + (T-234)²]^{1/2}. The 10 most similar base counts are shown.

Rank	Score	LOCUS	Definition	Base Count			
				A	C	G	T
1	0.0	HUMP5311*	human p53 tumor suppressor	276	364	308	234
2	1.4	HUMTP53A	human p53 tumor suppressor	276	365	307	234
3	2.8	HUMP53C	human p53 tumor suppressor	276	366	306	234
4	2.8	HUMTP53B	human p53 tumor suppressor	276	366	306	234
5	3.7	HUMP53T	human p53 tumor suppressor	277	366	305	234
6	15.7	MUSP53B	mouse p53 tumor suppressor	278	350	305	240
7	16.0	MUSP53A	mouse p53 tumor suppressor	279	350	304	240
8	16.3	HUMPAIA	human plasminogen activator inhibitor	275	374	316	244
9	16.3	HUMPAI	human plasminogen activator inhibitor	275	374	316	244
10	16.3	HUMPAIR	human plasminogen activator inhibitor	275	375	316	243

Table 4. Codon usage tables for zeta-globin genes from human (HUM), chimpanzee (CHP) and horse (HRS). Codon usage is from Wada *et al.* (3).

Codon	HUM	CHP	HRS	Codon	HUM	CHP	HRS	Codon	HUM	CHP	HRS	Codon	HUM	CHP	HRS
TTT Phe	0	0	0	TCT Ser	2	2	1	TAT Tyr	0	0	0	TGT Cys	0	0	1
TTC Phe	7	7	7	TCC Ser	7	7	8	TAC Tyr	3	3	4	TGC Cys	1	1	0
TTA Leu	0	0	0	TCA Ser	0	0	0	TAA Stop	0	0	0	TGA Stop	1	1	1
TTG Leu	1	1	0	TCG Ser	1	1	1	TAG Stop	0	0	0	TGG Trp	2	2	2
CTT Leu	0	1	0	CCT Pro	0	0	0	CAT His	0	0	0	CGT Arg	0	0	0
CTC Leu	3	2	2	CCC Pro	1	1	1	CAC His	7	7	6	CGC Arg	4	4	4
CTA Leu	1	1	0	CCA Pro	0	0	0	CAA Gln	0	0	0	CGA Arg	0	0	0
CTG Leu	12	12	14	CCG Pro	4	4	4	CAG Gln	3	3	4	CGG Arg	0	0	0
ATT Ile	1	1	0	ACT Thr	2	2	0	AAT Asn	0	0	0	AGT Ser	0	0	0
ATC Ile	6	6	4	ACC Thr	8	8	5	AAC Asn	1	2	2	AGC Ser	3	3	4
ATA Ile	0	0	1	ACA Thr	0	0	0	AAA Lys	0	0	0	AGA Arg	0	0	0
ATG Met	2	2	3	ACG Thr	2	2	3	AAG Lys	9	9	9	AGG Arg	2	1	2
GTT Val	0	0	0	GCT Ala	0	1	1	GAT Asp	0	0	1	GGT Gly	0	0	0
GTC Val	4	3	5	GCC Ala	12	11	12	GAC Asp	8	7	7	GGC Gly	5	5	6
GTA Val	1	1	0	GCA Ala	0	0	0	GAA Glu	0	0	0	GGA Gly	0	0	0
GTG Val	6	7	7	GCG Ala	4	4	6	GAG Glu	6	6	5	GGG Gly	1	2	0

Although the codon table does not define the order in which the codons appear and therefore the sequence of the gene, the chance that two unrelated sequences would have the same codon table is negligible. Consequently, codon usage may function to identify a sequence uniquely. Of course, identical coding sequences, such as those arising from recent gene duplication or conversion events, will have identical codon tables and cannot be distinguished on this basis. Older duplicated coding sequences may be differentiated even if the corresponding protein is perfectly conserved. For example, two human calmodulin genes, A₁₃₃C₈₉G₁₃₉T₈₆ and A₁₁₃C₁₁₀G₁₂₀T₁₀₄, code an identical amino acid sequence with a very different pattern of synonymous codon usage (19).

Homologous genes from different organisms can be expected to have different codon tables on account of nucleotide substitutions in the corresponding genes, especially those occurring in silent (degenerate) sites of codons. Coding sequences evolve primarily through the substitution of one base for another or through the deletion or insertion of triplets of bases. The deletion or addition of a single triplet causes one entry in the corresponding codon table to change. Each single nucleotide substitution, silent or replacement, alters two entries. Silent nucleotide substitutions typically occur in nuclear genes of vertebrates at a rate of about one per 200 sites per million years (20). Accordingly, even genes coding for a protein whose amino acid sequence has been perfectly conserved, may be easily distinguished on the basis of codon tables or base counts. Because there is a very large number of ways to code a typical protein using different numbers of

synonymous codons, the likelihood that orthologous genes should reconverge on the same codon table is slight.

The human zeta-globin gene was again compared to 9,037 sequences from GenBank 62.0, but this time using codon tables rather than base counts. Figure 1 summarizes the distribution of numbers of matches between the hemoglobin zeta-globin table and those of the other genes. The zeta-globin gene of chimpanzee is the only other codon table having more than 35/64 matches. The average number of matches is 10.3. If we assume, based on Figure 1, that the chance of a pair of unrelated genes having the same number of a particular codon is less than 0.2, then it follows that the likelihood of a pair of unrelated sequences having identical usage of all codons is less than $(0.2)^{64} = 1.8 \times 10^{-45}$.

Base counts and codon tables may serve as the basis of a gene index based on sequence that can be used to cross-reference other information such as GenBank DEFINITIONS and LOCUS names. Table 5 contains a partial listing of human coding sequences sorted by base count. Entries correspond to human

Figure 1. Comparison of human zeta-globin codon table with those of 9,037 genes from GenBank 62.0. A codon table is said to match the zeta-globin table when it has the same number of a particular codon, e.g. TTC₇ (Table 4). The numbers of genes having from 1 to 64 matches are shown. The average number of matches is 10.3. GenBank LOCUS names WHTH3, GOTAGLZ2, CHPAZGLO, HUMHBA1, and HUMXIIB represent wheat histone H3 (33 matches), goat zeta-globin (35 matches), chimp zeta-globin (54 matches), and human zeta-globin (two entries, 64 matches), respectively.

Table 5. Human gene index based on base counts. This is a partial listing of a directory of 1,490 complete reading frames for human genes from GenBank 62.0. Entries are listed in order of increasing A, C, G, T, respectively. Identical and nearly identical entries are marked with an asterisk. Base counts were derived from codon usage tables (3).

A	Base Count			GenBank LOCUS	No. Codons	Definition
C	G	T				
* 285	181	243	254	HUMBCI	321	blood coagulation inhibitor
* 285	181	243	254	HUMENN	321	endonexin II
* 285	181	243	254	HUMPAP4	321	placenta anticoagulant protein PP4
* 285	181	243	254	HUMLC5	321	lipocortin-V
* 285	181	243	254	HUMATC	321	placental anticoagulant protein (PAP)
* 285	181	243	254	HUMVAC	321	vascular anticoagulant
285	217	298	280	HUMPDHB	360	pyruvate dehydrogenase beta subunit
285	346	297	230	HUMLMGP	386	lysosome-associated membrane glycoprotein
285	443	458	296	HUMPEPD	494	prolidase
285	488	425	287	HUMMHCP42	495	steroid-21 hydroxylase B
* 286	195	179	234	HUMBILYM	298	B-lymphocyte cell surface antigen B1
* 286	196	179	233	HUMCD20A	298	CD20 receptor
* 286	197	179	232	HUMCD20	298	B-lymphocyte antigen CD20
286	349	314	272	HUMRENA5*	407	renin
287	266	274	235	HUMHODB3*	354	3-beta hydroxysteroid dehydrogenase
* 287	418	440	241	HUMPRCA	462	protein C
* 287	418	441	240	HUMPRCM	462	protein C
* 287	420	440	242	HUMPRC7*	463	protein C
288	165	244	302	HUMLDHX	333	lactate dehydrogenase
288	241	279	236	HUMHPA1S	348	haptoglobin alpha-1S
288	427	369	215	HUMRAR	433	retinoic acid receptor

genes from GenBank 62.0 with A₂₈₅ to A₂₈₈ inclusive.. Such an index, ordered by base counts or codon tables, can be used like a telephone directory to help locate similar database entries or searched rapidly using simple computer algorithms.

Finding Homologous Genes

In addition to providing a 'fingerprint' to identify sequences (Figure 1), information in the codon table may be used to infer evolutionary relationships. The advantage of this approach is that it does not require sequence alignment and large amounts of computation. We have used the following difference formula to compare codon tables:

$$\text{Score} = [(L_1 - L_2)^2 + \sum_{i=1}^{20} (N_{1i} - N_{2i})^2]^{1/2}$$

where N_{1i} and N_{2i} represent the counts of synonymous codons for each of the 20 amino acids and L_1 and L_2 represent the total number of codons in two sequences represented by codon tables. Additionally, the terms in the sum corresponding to Cys and Trp are weighted by 5 and those corresponding to Gly, Pro, Phe, Leu, and Tyr are weighted by 2 because these amino acids typically undergo replacement less frequently than the others (21). This measure, like those proposed by Cornish-Bowden (22,23), relies on amino acid composition to predict sequence similarity.

As an example, the codon table for the human green photoreceptor pigment sequence (24) was compared to those of 9,037 sequences from GenBank 62.0 using the above formula. The most similar scores are listed in Table 6. The recently diverged (30 MYA, 24) red cone pigment is readily distinguished from unrelated sequences and is followed by several homologous opsins and G-coupled proteins. The blue cone pigment protein, ranked sixth in similarity, has 44% amino acid identity and is estimated to have diverged at least 500 MYA (24).

Table 6. Similarity search based on codon tables. The codon table for human green cone photoreceptor (HUMCNPG6) was compared to 9,037 others from GenBank 62.0 using the difference measure in the text. The lowest scores are shown.

Rank	Score	LOCUS	Definition
1	0.0	HUMCNPG6*	human green cone photoreceptor pigment
2	4.7	HUMCNPR6*	human red cone photoreceptor pigment
3	30.2	DROPSA	<i>Drosophila</i> opsin
4	30.5	DRORH4A2*	<i>Drosophila</i> opsin RH4
5	31.2	DROOPSB5*	<i>Drosophila</i> opsin
6	32.1	HUMCNPB5*	human blue cone photoreceptor pigment
7	32.2	DROOPSAA	<i>Drosophila</i> opsin RH2
8	34.2	BOVSKR	bovine substance K receptor
9	35.2	RATMPT	rat mt proton/phosphate symporter
10	35.6	FDIPSBA	cyanobacterium herbicide binding
11	36.2	ANIPSBA2	<i>Anacystis</i> photosystem II
12	36.2	ANIPSBA3	<i>Anacystis</i> photosystem II
13	36.4	BOVPCR	bovine mito. phosphate carrier
14	36.4	HUMOPS	human opsin
15	36.9	ANIPSBA	<i>Anacystis</i> psbA, photosystem
16	36.9	ANIPSBA1	<i>Anacystis</i> R2 psbA1, photosystem II
17	37.0	RATCONN	rat connexin 43
18	37.2	DRORH3A	<i>Drosophila</i> opsin RH3
19	37.2	DRORH92CD	<i>Drosophila</i> opsin RH7
20	38.2	DOGGPCR1	G-protein coupled receptor
21	38.3	CHKLKPA	chicken cartilage link protein
22	38.4	BOVOPS5*	bovine opsin
23	38.6	SYCSBA1	Synechocystis 6903 psbA-1 photosystem
24	38.7	BOVOBCAM	bovine opioid binding protein
25	39.4	EGRCPSPB	<i>Euglena gracilis</i> chloroplast psbA

Table 7 shows the results of a similar comparison of the human zeta-globin codon table (Table 4) with the others. Again, only the lowest scores, resulting from the closest similarities, are shown. The six most similar are the zeta-globins of various organisms and these are followed by 52 other globins until the first non-globin (human prealbumin) is reached. The amino acid composition and size of human prealbumin are similar to globins.

Because it is computationally simple, this method is relatively fast. The searches shown in Tables 6 and 7 each require less than 15 seconds on a microcomputer. This technique, based on global comparisons, is necessarily limited to complete coding sequences, however.

Identifying Loci

It may be possible to designate protein-coding genes with base counts and codon tables, although there are several difficulties with this approach. More than one sequence, and therefore multiple designations, may accrue to a particular locus because of population variation or sequencing errors (e.g. Table 3). Accordingly, a convention would have to be devised for deciding which sequence to use as the standard and symbols would have to be revised if a reported coding sequence was found to be in error. Another circumstance is that, because of gene duplication, the same sequence may exist at multiple loci. These would have to be distinguished by considering additional characters such as chromosomal location.

Table 7. Similarity search based on codon tables. The codon table for human zeta-globin (HUMHBA1) was compared to 9,037 others from GenBank 62.0 using the difference measure in the text. The lowest scores are shown.

Rank	Score	LOCUS	Definition
1	0.0	HUMHBA1	human zeta-globin
2	0.0	HUMXIHB	human zeta-globin
3	2.2	CHPAZGLO	chimpanzee zeta globin
4	5.5	GOTAGLZ2*	goat zeta-globin
5	6.3	HRSZGL1	horse zeta-globin
6	6.3	HRSZGL2	horse zeta-globin
7	8.2	DUKHGAP	duck embryonic alpha-globin pi
8	9.7	XELHBAT4X	<i>Xenopus laevis</i> tadpole alpha-globin
9	10.0	XETGLA	<i>Xenopus tropicalis</i> alpha-globin
10-47	10.0-14.1	(alpha- and beta-globins, various organisms)	
48	14.2	SOYLBGI	soybean leghemoglobin i
49	14.2	XELHBA2M	<i>Xenopus laevis</i> alpha-2-globin
50	14.6	XELHBB1	<i>Xenopus laevis</i> beta-1-globin
51	14.7	MACHBCA2	rhesus monkey gamma-globin
52	14.7	MACHBGA1	rhesus monkey gamma-globin
53	14.8	XETHBBA	<i>Xenopus tropicalis</i> alpha-globin
54	14.8	MUSHBBH1	mouse embryonic beta-globin H1
55	14.8	CHPGGGLOG	chimpanzee G-gamma-globin
56	14.8	HUMHBB # 2	human beta-globin
57	14.8	ORAHBG1F	orangutan gamma-1-globin
58	14.9	DUKHBADB	duck alpha-d globin
59	14.9	DUKHBADWP	duck alpha-II globin
60	14.9	BOVHBBE2	bovine epsilon-2 beta-globin
61	14.9	HUMHBB # 1	human beta globin
62	15.1	HUMPALF3*	human prealbumin
63	15.1	HUMPALFAP	human prealbumin
64	15.1	CHKHBBR1	chicken rho-globin
65	15.1	CHKHBRHO2	chicken rho'-globin
66	15.2	HUMPALA	human prealbumin
67	15.2	HUMPALB	human prealbumin
68	15.2	HUMPALD	human prealbumin
69	15.2	ORAHBG2F	orangutan gamma-2-globin
70	15.2	XELHBBAI	<i>Xenopus laevis</i> alpha(I)-globin

Limitations of numerical classification

A disadvantage of the methods we have discussed is that they require the coding sequence of a gene to have been completely and accurately determined; partial coding sequences cannot be described. If the sequence or boundaries of a reported coding region are later found to be in error, the corresponding base count and codon table would also have to be revised. Nucleotide sequences that differ only in non-coding regions (or in chromosomal location) will necessarily have the same description and cannot be differentiated solely by these means. Finally, numerical descriptions are not evocative, unlike mnemonics that are suggestive of function.

CONCLUSION

We have described simple methods for characterizing the nucleotide sequences of genes. Numerical classification of coding sequences is useful for two reasons. First, sequence-based designations, such as base counts and codon tables, do not require revision as knowledge of function evolves. Accordingly, these descriptors are well-suited to cross-reference entries within or between databases. Second, sequence-based annotation provides information that can be used to infer evolutionary relationships, without the need to align actual nucleotide sequences. These methods may be used to complement the annotation used in GenBank and elsewhere.

ACKNOWLEDGMENTS

We thank Dr. Toshimichi Ikemura for kindly providing us with a magnetic tape containing codon usage data from GenBank and two anonymous reviewers whose comments helped to clarify the discussion. This work was supported by NIH grant HG00312 to the University of California, Berkeley.

REFERENCES

1. Bilofsky, H.S. and Burks, C. (1988) *Nucleic Acids Res.*, **16**, 1861–1864.
2. Release notes, GenBank 69.0.
3. Lauer, K., Shen, J. and Maniatis, T. (1980) *Cell*, **20**, 119–130.
4. Shows, T.B., McAlpine, P.J., Boucheix, C., Collins, F.S., Conneally, P.M., Frezal, J., Gershowitz, H., Goodfellow, P.N., Hall, J.G., Issit, P., Jones, C.A., Knowles, B.B., Lewis, M., McKusick, V.A., Meisler, M., Morton, N.E., Rubenstein, P., Schanfield, M.S., Schmickel, R.D., Skolnick, M.H., Spence, M.A., Sutherland, G.R., Traver, M., Van Cong, N. and Willard, H.F. (1987) *Cytogenet. Cell. Genet.*, **46**, 11–28.
5. De Jong, W.W. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 7114–7118.
6. Levine, A.J., Momand, J., and Finlay, C.A. (1991) *Nature*, **351**, 453–456.
7. Nusse, R., Brown, A., Papkoff, J., Scambler, P., Shackleford, G., McMahon, A., Moon, R., Varmus, H. (1991) *Cell*, **64**, 231–232.
8. Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F., Kerlavage, A.R., McCombie, W.R. and Venter, J.G. (1991) *Science*, **252**, 1651–1656.
9. Doolittle, R.F. (1990) In Bell, G. (ed.), *Computers and DNA*. (SFI Studies in the Sciences of Complexity, vol VII). Addison-Wesley, pp. 21–31.
10. Wada, K., Aota, S., Tsuchiya, R., Ishibashi, F., Gojobori, T. and Ikemura, T. (1990) *Nucleic Acids Res.*, **18**, Supplement, 2367–2411.
11. Funakoshi, T., Heimark, R.L., Hendrickson, L.E., McMullen, B.A., and Fujikawa, K. (1987) *Biochemistry*, **26**, 8087–8092.
12. Iwasaki, A., Suda, M., Nakao, H., Nagoya, T., Saino, Y., Arai, K., Shidara, Y., Murata, M., and Maki, M. (1987) *J. Biochem.*, **102**, 1261–1273.
13. Kaplan, R., Jaye, M., Burgess, W.H., Schlaepfer, D.D., Haigler, H.T. (1988) *J. Biol. Chem.*, **263**, 8037–8043.
14. Pepinsky, R.B., Tizard, R., Mattaliano, J., Sinclair, L.K., Miller, G.T., Browning, J.L., Chow, P., Burne, C., Huang, K., Pratt, D., Wachter, L., Hession, C., Frey, A. and Wallner, B.P. (1988) *J. Biol. Chem.*, **263**, 10799–10811.
15. Grundmann, U., Abel, K., Bohn, H., Lbermann, H., Lottspeich, F. and Kpper, H. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 3709–3712.
16. Maurer-Fogy, I., Reutelingperger, C.P.M., Pieters, J., Bodo, G., Stratowa, C. and Hauptman, R. (1988) *Eur. J. Biochem.*, **174**, 585–592.
17. Rojas, E., Pollard, H.B., Haigler, H.T., Parra, C. and Burns, A.L. (1990) *J. Biol. Chem.*, **265**, 21207–21215.
18. Matlashewski, G., Lamb, P., Pim, D., Peacock, J., Crawford, L. and Benchimol, S. (1984) *EMBO J.*, **3**, 3257–3262.
19. Fischer, R., Koller, M., Flura, M., Mathews, S., Strehler-Page, M.A., Krebs, J., Penniston, J.T., Carafoli, E. and Strehler, E.E. (1988) *J. Biol. Chem.*, **263**, 17055–17062.
20. Li, W.H., Chung, I.W., Luo, C.C. (1985) *Mol. Biol. Evol.*, **2**(2), 150–174.
21. Dayhoff, M.O. (1978) *Atlas of Protein Sequence and Structure* (National Biomedical Research Foundation, Silver Spring, Md.), vol. 5, supplement 3.
22. Cornish-Bowden, A. (1977) *J. Theor. Biol.*, **65**, 735–742.
23. Cornish-Bowden, A. (1979) *J. Theor. Biol.*, **76**, 369–386.
24. Nathans, J., Thomas, J. and Hogness, D.S. (1986) *Science*, **232**, 193–202.