# The majority of human genes have regions repeated in other human genes

**Roy J. Britten***

California Institute of Technology, 101 Dahlia Avenue, Corona del Mar, CA 92625

Contributed by Roy J. Britten, February 17, 2005

**Amino acid sequence comparisons have been made between all of 25,193 human proteins with each of the others by using BLAST software (National Center for Biotechnology Information) and recording the results for regions that are significantly related in sequence, that is, have an expectation of $<1 \times 10^{-3}$. The results are presented for each amino acid as the number of identical or similar amino acids matched in these aligned regions. This approach avoids summing or dealing directly with the different regions of any one protein that are often related to different numbers and types of other proteins. The results are presented graphically for a sample of 140 proteins. Relationships are not observed for 26.5% of the 12,728,866 amino acids. The average number of related amino acids is 36.5 for the majority (73.5%) that show relationships. The median number of recognized relationships is ≈3 for all of the amino acids, and the maximum number is 718. The results demonstrate the overwhelming importance of gene regional duplication forming families of proteins with related domains and show the variety of the resulting patterns of relationship. The magnitude of the set of relationships leads to the conclusion that the principal process by which new gene functions arise has been by making use of preexisting genes.**

domains | protein | relationships

It has long been clear that genes come in families, with selection for functionally useful domains. Genes can be thought to be the result of past events of domain shuffling, splicing, fusion, deletion, and duplication during the evolution of protein families. This article reports the results of an attempt to count the amino acid sequence relationships among all of the presently identified human genes. It amounts to an overview. For this purpose, the relationships of individual amino acids are counted when they are part of regions of proteins that have significant relationships to other proteins. We find that the vast majority of the proteins are involved in such relationships and on average are related to ≈26 other proteins.

## Methods

The 25,193 human gene amino acid sequences were downloaded from the National Center for Biotechnology Information. Using BLAST, each sequence as probe was compared with all of the other sequences (called targets), rejecting any alignments with an expectation value $>1 \times 10^{-3}$. In comparing amino acid sequence pairs, BLAST often reports alignments of short regions of particular probes and targets that may multiply overlap. Each amino acid of the probe that was matched to either an identical or similar amino acid within one of these regions was given a score of 1, regardless of how many times it was reported in overlapping regions with the same target. Then, the scores of all of the matches to different targets were added up to determine for each amino acid in each gene the number of matches present within regions of significant relationships. A table for the 25,193 genes was constructed in which a name index and the length are listed for each, followed by a listing of the number of similar or identical amino acids recognized for each of the amino acids of the protein. This 30-megabyte table is available on CD by e-mail request. Also available on CD is FORTRAN software which allows the graphical presentation for each gene of the patterns of frequency of matches for all amino acids. The identification number in Table 2 and Table 3, which is published as supporting information on the PNAS web site, is derived from the National Center for Biotechnology Information code NM_xxxxxx for the mRNA for that protein, which is converted to 1xxxxx. For the CD disk, the identification method is to search the PubMed nucleotides database for "*Homo sapiens*" and a protein name or identifier. Among the results will be the NM_xxxxxx, which is converted as above to find the protein on the disk, or XM_xxxxxx, which is converted to 2xxxxxx.

## Results

**Numbers of Matches.** The approach counts all of the individual amino acid matches (identical or similar) that are part of a region of a gene that has significant sequence relationships with other genes. In Table 1, the two columns show percent matches with an expectation value of $<1 \times 10^{-3}$ to the left and $<1 \times 10^{-6}$ to the right. There are no important differences between the two columns, and an expectation $<1 \times 10^{-3}$ was considered significant.

Genes often include significant matches to different sets of other genes in different regions of the gene, making it difficult to count the number of relationships of a gene as a whole. Therefore, a direct and simple method for counting the number of individual amino acid relationships is preferable, avoiding counting the sum or average of all of the relationships of different regions. This analysis was done with 25,193 genes that on average each contain 505 aa, adding to a total length of 12,728,886 aa. There were 339 million matches between amino acids in regions with an expectation $\leq 1 \times 10^{-3}$, giving an average of 36.5 matches per amino acid, counting just those amino acids with any matches. Table 1 summarizes the number of matches for the individual amino acids of each of the human genes in this fairly complete set with the other human genes in this set. The data are restricted to significantly matching regions with expectations of $1 \times 10^{-3}$.

Of interest is the number of genes that do or do not contain matching sequences. By count, 4,286 genes, or only 17.0% of the total of 25,193 genes, do not have any matching regions with expectation $\leq 1 \times 10^{-3}$. By whatever measure is chosen, the majority of genes or amino acids of gene sequences are part of sets of related sequences. Seventy-three percent of the amino acids match others in regions of significant relationship, with an average of 36.5 matches for amino acids in genes with regions that match significantly. The number of matches ranges from 0 to 718 for given amino acids. The highest-frequency amino acids are part of C2H2 zinc finger proteins.

Although the average number of matches for all of the amino acids of all genes on this list is ≈26, the median is nearly 3. One view of this observation is that the majority of amino acids are in regions of proteins significantly related to few other proteins

Freely available online through the PNAS open access option.

*E-mail: r.britten@comcast.net or rbritten@caltech.edu.





www.pnas.org/cgi/doi/10.1073/pnas.0501008102

**Table 1. Individual matches for 12,728,866 amino acids**

| BLAST expectation | $1.0 \times 10^{-3}$ | $1.0 \times 10^{-6}$ |
|---|---|---|
| % amino acids matched | 72.86 | 70.51 |
| % one match* | 14.11 | 14.67 |
| % genes with match† | 83.25 | 80.51 |
| Average no. of matches per amino acid | 26.6 | 23.5 |
| Average no. of matches per matched amino acid | 36.5 | 33.33 |

*Amino acids with one match are matched in a region that is recognizably present in just two different proteins.

†Percent of proteins that contain matching regions with an expectation value $<1 \times 10^{-3}$ or $<1 \times 10^{-6}$.

and, as a result, have none, one, two, or three matches. The other view of this value of the median number of matches is that about half of the amino acids are in regions significantly related to many other proteins, ranging from 1 to 718 matches.

The arrangement of these amino acid similarities on the genes is important because they each are part of regions of the gene sharing functional similarity with other genes. Fig. 1 is an example of a graphical method showing the count of relationships of the amino acids of zinc finger gene 91, which is of C2H2 type. The maximum number of similarities for a given amino acid is 692 in this gene, and the gene length is 1,191 aa. In Fig. 1, the number of matches for each amino acid is plotted against position along the length of the protein sequence. The axes of the figure have been normalized so that the maximum frequency is near the top and the length fits the longitudinal size of the figure to set a standard way of exhibiting many cases. A feature of this example occurs widely among the set of genes studied. Each region containing matches is a mixture of amino acids showing lower frequencies and amino acids with a large number of matches. This range makes for a large vertical spread to the diagrams. In the case of Fig. 1, there is a long, level region of a typical maximum number of matches, and almost the whole gene has a high frequency of matches. This gene is made up of many repeats of the active zinc finger domain.

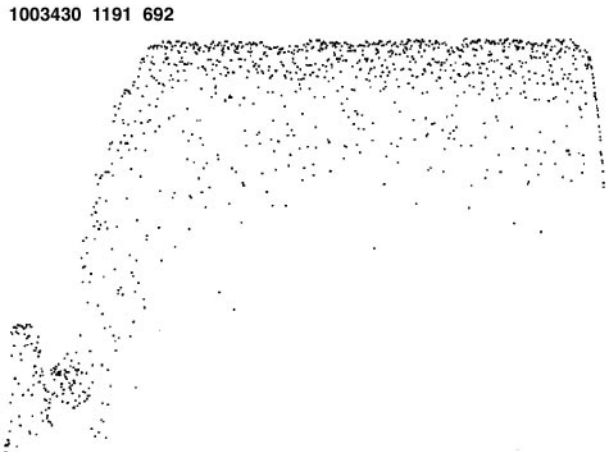


**Fig. 1.** Profile of frequency of match for zinc finger gene 91, which is a C2H2 type. The number of matches of individual amino acids in regions significantly matching other proteins in a collection of 25,193 human proteins is plotted against position in the ZF91 protein. The axes have been normalized so that the image just fits the space allotted to the figure. (*Upper Left*) As shown by the numbers, the length of ZF91 is 1,191 aa, and the maximum frequency is 692. Two amino acids at the left or NH4 terminal have few recognized copies (8 and 9) in significantly related regions, effectively supplying the zero position for the plot.

Fig. 2 shows similar diagrams for a sample of 70 different genes. Table 2 identifies the proteins shown in Fig. 2, listing also the length and maximum frequency. Fig. 4 and Table 3, which are published as supporting information on the PNAS web site, list the next 70 cases as shown in Fig. 2 and Table 2. Listed on these tables is a sample of 165 proteins chosen simply as the first 165 on the list with which we are working. (Note that the starting point is not 1, and that three are missing from the serial order.) Of this subset, 25 did not have any amino acid matches recognized as part of regions with an expectation of $\leq 1 \times 10^{-3}$. That amounts to 15.2%, which is not significantly different from the value of 17.0% for the whole set shown in Table 1, suggesting that this may be a typical sample. This set has not been purposely selected in any way, but it has the advantage that no computer-derived apparent genes are present. The 140 cases in which there are matches are illustrated in Figs. 2 and 4. Locations in these figures are identified by digits in rows 1–7 (vertical) and in columns 0–9 (horizontal). All examples are normalized by frequency and length, as in Fig. 1, to fit their individual boxes in the figure. For example, case 1,0 (row, column) at the upper left in Fig. 2 reports a duplicated gene where one copy has diverged from the other at a modest number of amino acids scattered along its length. PubMed lists these two genes as NAT1 and NAT2. Case 2,2 reports, for example, five essentially identical genes, which are apparently transcript variants, indicating what could be considered a source of minor statistical error in this work due to transcription variants being counted as genes.

In further introductory explanation, just looking at the cases in the figure usually allows recognition of the frequencies by the number of lines in the figure. These lines represent lower numbers of matches for the differing amino acids in the various related genes. The first five cases in the top row of Fig. 2 are examples where the low frequency can be easily recognized. However, case 2,2 is a rare example where the frequency cannot be recognized from the graph and must be derived from Table 2. The digestion of all of the immense amount of data is left up to the reader, depending on special interests. There are all kinds of examples ranging from genes that show similarity to other genes only in limited very small regions to those that are similar to many others over most of their length. There are cases of two distinct regions related to many others and even multiple regions, separated by regions with smaller numbers of matches. The six examples at the bottom right in Fig. 1 are a set of collagens that share a similar pattern of high frequency over most of their lengths. The maximum frequency of 718 observed is for an amino acid in the conserved region of a zinc finger protein and other high-frequency examples are from C2H2 zinc finger proteins.

Fig. 3 shows the distribution against the frequency of the number of amino acids on a log scale because it covers such a wide range. The distribution is expressed as the sum of the number of examples up to each observed frequency. The 27% of the amino acids that do not match others in significant regions start off the curve because there are 3.45 million of them. There are more with this value than for any other individual frequency. The upturn at frequencies >512 is very likely due to the conserved amino acids of the C2H2 zinc finger proteins, which adds up to a couple percent of all of the genes studied.

A question arises concerning the significance of the estimate of the number of genes without regions matching other genes, the 17% that are presumptive single-copy genes. To check this observation, the protein sequences that had no matches at an expectation of $1 \times 10^{-3}$ were collected in a library and compared with all 25,193 protein amino acid sequences at a more open criterion listing all relationships with an expectation <10. This same library was randomized in amino acid sequence, retaining the length and composition of each of the 4,180 proteins. The randomized set also was compared with all 25,193 protein

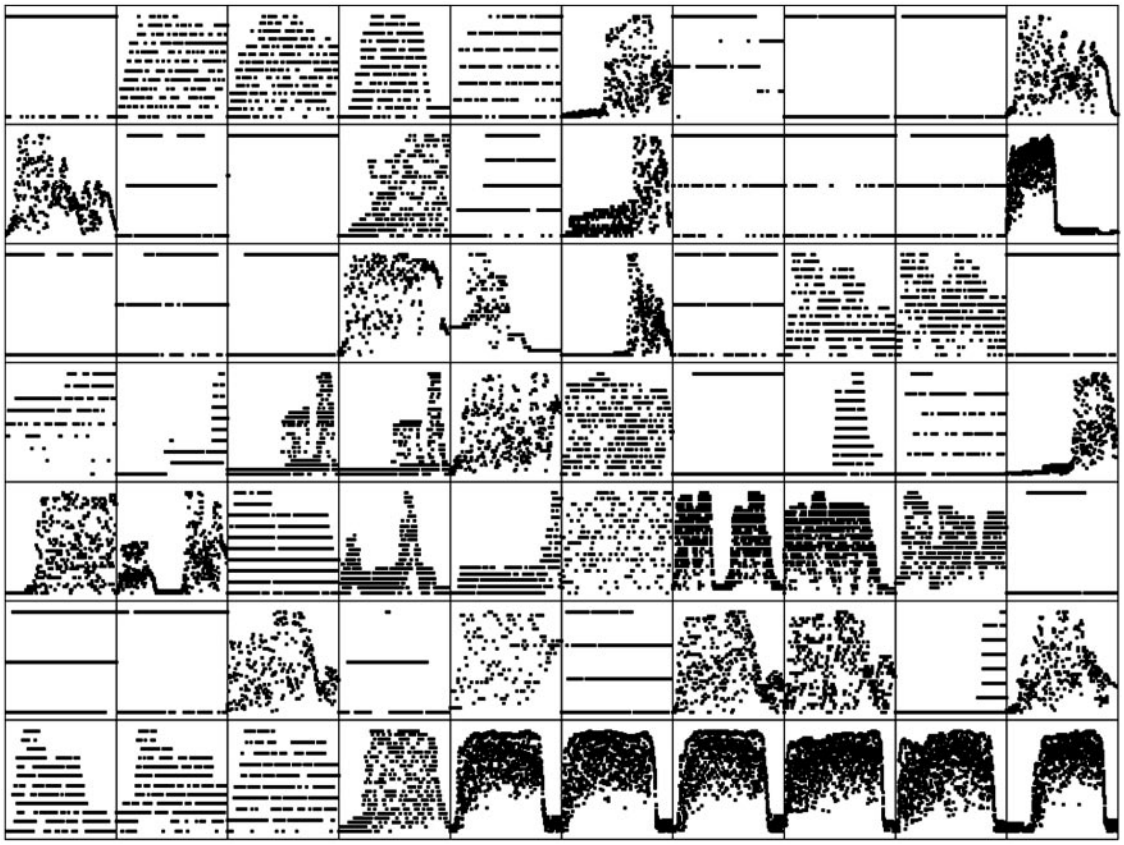

**Fig. 2.** Profiles for 70 human proteins. Shown are the first 70 cases in which there are matches. Table 2 identifies the individual proteins, and their locations are identified in columns 4 and 5 by numbers 1–7 (Y) for the rows (vertical) and 0–9 (X) for the columns (horizontal). All examples are normalized by frequency and length, as in Fig. 1, to fit their individual boxes on the figure. For example, case 1,0 (row, column) (*Upper Left*) reports a duplicated gene where one copy has diverged from the other at a modest number of amino acids scattered along its length, as shown by the dotted line at zero. Table 2 also lists the length of the protein and maximum frequency of any amino acid in the protein.

sequences. The maximum number of matching genes with expectation <10 was 252 for the original and 31 for the randomized set. Curiously, all of the randomized set had at least one match, whereas 190 of the original set found no matches at this open criterion.

To estimate how many of this set of genes can be considered to have significant matches with expectation $>1 \times 10^{-3}$, note that 85 of them have more matches than any of the randomized examples. This small number is hardly worth considering and suggests that the limit of $1 \times 10^{-3}$ is reasonable. Because of evolutionary relationships, many of the so-called single-copy proteins may, in fact, have domains that are very divergently related to the well recognized functional domains. It will be difficult to determine the number in this set because of the well recognized problem of seeing distant relationships. Each member of the set of 26,193 was retested against all of the set by using a new version of BLAST, and the results were consistent with those shown in Figs. 2 and 4. Graphs show the same general shape in every case but appear to include more "hits," indicating an increase in sensitivity to more divergent relationships, even though the limit was set to the same value of $1 \times 10^{-3}$. The percentage showing no relationships was 15.5%, compared with the previous 17%. Results with either BLAST version would be satisfactory, but the distributed CD contains the newer, more sensitive version results. This paper can be considered just the beginning of this sort of study of relationships with the tools at hand. Future studies, for example, by setting up hidden Markov model analysis, will likely show many additional relationships.

**A Translated Fragment of an Alu Repeat in a Zinc Finger Gene.** When REPEATMASKER (Institute for Systems Biology, Seattle) software examined the coding sequence of the zinc finger gene 91 described in Fig. 1, it reported that the last 104 nucleotides (or ≈34 aa) are part of an Alu Ye sequence matching at 94%. It turns out that the last exon of this gene is short and is made up only of the fragment of Alu sequence, similar to Alu Ye. Thus, this is probably an example of a gene variation caused by the presence of an exon derived from a mobile element (1–3). A search turned up five other cases of zinc finger genes that have similar short terminal Alu sequences, as well as a few more zinc finger genes with fragments of other mobile elements inserted into the coding sequences.

Regarding the interpretation of Fig. 1, the smoothly downsloping region at the right-hand end is mostly the Alu sequence translation. This region indicates that there are hundreds of related amino acid sequence regions that include these amino acids, but it does not show that they are linked to zinc finger genes. When REPEATMASKER is used to search all of the 25,193 gene coding sequences for mobile elements, REPEATMASKER finds many hundreds but adding up to <1% of the length of these coding regions. Among them are ≈74 examples of fragments of Alu sequences in genes that produce known proteins. Presumably, some of these Alu sequences are part of the sequences that are matched by the amino acids of the terminal short region of Fig. 1, and they occur among many different gene coding sequences. However, the short terminal region that descends like the tail of an elephant at the right edge of Fig. 1 indicates that there are >400 similar amino acids in significantly related

**Table 2. List of proteins in Fig. 2**

| ID | Length | Frequency | Y | X | Description |
|---|---|---|---|---|---|
| 1000015 | 290 | 1 | 1 | 0 | N-acetyltransferase 2 (NAT2) |
| 1000016 | 421 | 11 | 1 | 1 | Acyl-Coenzyme A dehydrogenase, (ACADM) |
| 1000017 | 412 | 13 | 1 | 2 | Acyl-Coenzyme A dehydrogenase, (ACADS) |
| 1000018 | 655 | 12 | 1 | 3 | Acyl-Coenzyme A dehydrogenase, (ACADVL) |
| 1000019 | 427 | 6 | 1 | 4 | Acetyl-Coenzyme A acetyltransferase 1 (ACAT1) |
| 1000020 | 503 | 241 | 1 | 5 | Activin A receptor type II-like 1 (ACVRL1) |
| 1000021 | 467 | 4 | 1 | 6 | Presenilin 1 (Alzheimer disease 3) (PSEN1) |
| 1000022 | 363 | 1 | 1 | 7 | Adenosine deaminase (ADA) |
| 1000023 | 387 | 1 | 1 | 8 | Sarcoglycan, alpha (SGCA) |
| 1000024 | 413 | 230 | 1 | 9 | Adrenergic, beta-2-, receptor, surface (ADRB2) |
| 1000025 | 408 | 197 | 2 | 0 | Adrenergic, beta-3-, receptor (ADRB3) |
| 1000027 | 346 | 2 | 2 | 1 | Aspartylglucosaminidase (AGA) |
| 1000028 | 1,532 | 5 | 2 | 2 | Amylo-1, 6-glucosidase, 4-alpha-glucanotrans. (AGL) |
| 1000029 | 485 | 19 | 2 | 3 | Angiotensinogen (serine (or cysteine) (AGT) |
| 1000032 | 587 | 4 | 2 | 4 | Aminolevulinate, delta-, synthase (ALAS2) |
| 1000033 | 745 | 32 | 2 | 5 | ATP-binding cassette, sub-family D (ALD) |
| 1000034 | 364 | 2 | 2 | 6 | Aldolase A, fructose-bisphosphate (ALDOA) |
| 1000035 | 364 | 2 | 2 | 7 | Aldolase B, fructose-bisphosphate (ALDOB) |
| 1000036 | 747 | 2 | 2 | 8 | Adenosine monophosphate deaminase 1 (AMPD1) |
| 1000037 | 1,880 | 171 | 2 | 9 | Ankyrin 1, erythrocytic (ANK1) |
| 1000038 | 2,843 | 1 | 3 | 0 | Adenomatosis polyposis coli (APC) |
| 1000039 | 267 | 2 | 3 | 1 | Apolipoprotein A-I (APOA1) |
| 1000041 | 317 | 1 | 3 | 2 | Apolipoprotein E (APOE) |
| 1000042 | 345 | 34 | 3 | 3 | Apolipoprotein H (beta-2-glycoprotein I) (APOH) |
| 1000043 | 335 | 26 | 3 | 4 | Tumor necrosis factor receptor (TNFRSF6) |
| 1000044 | 920 | 67 | 3 | 5 | Androgen receptor (AR) |
| 1000045 | 322 | 2 | 3 | 6 | Arginase, liver (ARG1) |
| 1000046 | 533 | 13 | 3 | 7 | Arylsulfatase B (ARSB) |
| 1000047 | 589 | 14 | 3 | 8 | Arylsulfatase E (ARSE) |
| 1000049 | 313 | 1 | 3 | 9 | Aspartoacylase (aminoacylase 2) (ASPA) |
| 1000050 | 412 | 8 | 4 | 0 | Argininosuccinate synthetase (ASS) |
| 1000051 | 3,056 | 9 | 4 | 1 | Ataxia telangiectasia mutated (ATM) |
| 1000052 | 1,500 | 23 | 4 | 2 | ATPase, Cu++ transporting, alpha polypep (ATP7A) |
| 1000053 | 1,465 | 25 | 4 | 3 | ATPase, Cu++ transporting, beta polypep (ATP7B) |
| 1000054 | 371 | 69 | 4 | 4 | Arginine vasopressin receptor 2 (AVPR2) |
| 1000055 | 602 | 17 | 4 | 5 | Butyrylcholinesterase (BCHE) |
| 1000056 | 392 | 1 | 4 | 6 | Branched chain keto acid dehydrog. E1 (BCKDHB) |
| 1000057 | 1,417 | 11 | 4 | 7 | Bloom syndrome (BLM) |
| 1000060 | 543 | 5 | 4 | 8 | Biotinidase (BTD) |
| 1000061 | 659 | 485 | 4 | 9 | Bruton agammaglobulinemia tyrosine kinase (BTK) |
| 1000062 | 500 | 34 | 5 | 0 | Serine proteinase inhibitor (SERPING1) |
| 1000063 | 752 | 65 | 5 | 1 | Complement component 2 (C2) |
| 1000064 | 1,663 | 9 | 5 | 2 | Complement component 3 (C3) |
| 1000065 | 934 | 21 | 5 | 3 | Complement component 6 (C6) |
| 1000066 | 591 | 16 | 5 | 4 | Complement component 8, beta polypeptide (C8B) |
| 1000067 | 260 | 18 | 5 | 5 | Carbonic anhydrase II (CA2) |
| 1000068 | 2,261 | 27 | 5 | 6 | Calcium channel, voltage-dependent, (CACNA1A) |
| 1000069 | 1,873 | 26 | 5 | 7 | Calcium channel, voltage-dependent, (CACNA1S) |
| 1000070 | 821 | 20 | 5 | 8 | Calpain 3, (p94) (CAPN3) |
| 1000071 | 551 | 1 | 5 | 9 | Cystathionine-beta-synthase (CBS) |
| 1000072 | 472 | 2 | 6 | 0 | CD36 antigen (collagen type I receptor) (CD36) |
| 1000073 | 182 | 1 | 6 | 1 | CD3G antigen, gamma polypeptide (CD3G) |
| 1000075 | 303 | 534 | 6 | 2 | Cyclin-dependent kinase 4 (CDK4) |
| 1000076 | 316 | 2 | 6 | 3 | Cyclin-dependent kinase inhibitor 1C (CDKN1C) |
| 1000077 | 156 | 24 | 6 | 4 | Cyclin-dependent kinase inhibitor 2A (CDKN2A) |
| 1000078 | 493 | 3 | 6 | 5 | Cholesteryl ester transfer protein, plasma (CETP) |
| 1000079 | 457 | 45 | 6 | 6 | Cholinergic receptor, nicotinic, alpha (CHRNA1) |
| 1000080 | 493 | 29 | 6 | 7 | Cholinergic receptor, nicotinic, (CHRNE) |
| 1000081 | 3,801 | 7 | 6 | 8 | Chediak-Higashi syndrome 1 (CHS1) |
| 1000082 | 396 | 53 | 6 | 9 | Cockayne syndrome 1 (classical) (CKN1) |
| 1000083 | 988 | 11 | 7 | 0 | Chloride channel 1, skeletal muscle (CLCN1) |
| 1000084 | 746 | 11 | 7 | 1 | Chloride channel 5 (CLCN5) |
| 1000085 | 687 | 9 | 7 | 2 | Chloride channel Kb (CLCNKB) |
| 1000087 | 690 | 23 | 7 | 3 | Cyclic nucleotide gated channel alpha 1 (CNGA1) |
| 1000088 | 1,464 | 103 | 7 | 4 | Collagen, type I, alpha 1 (COL1A1) |
| 1000089 | 1,366 | 102 | 7 | 5 | Collagen, type I, alpha 2 (COL1A2) |

Y is the row count with the top row = 1, and X is the column count with the left column = 0. For identification, see *Methods.*

regions. The interpretation of this region is not as clear and certain as we might prefer. It depends on the effect of the neighboring well matched regions on this region, the heuristic method used by BLAST to detect weakly similar regions, and the way the expectation is calculated to set the criterion of significance. That was demonstrated by using BLAST to compare

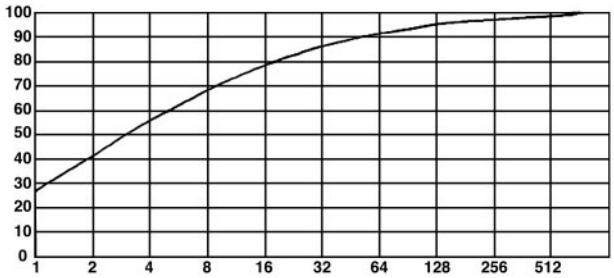


**Fig. 3.** Cumulative distribution of numbers of matches. The cumulated percent of amino acids plotted against the log of the number of identical or similar amino acids recognized, always restricted to significantly related aligned regions (expectation $<1 \times 10^{-3}$). The curve starts with the 27% of amino acids that have no matches, passes through the median at ≈3 matches, and passes through 90% with fewer than ≈50 matches. Finally, there is a step representing the zinc finger proteins and a few other proteins with between 500 and 718 matches.

terminal region segments of various lengths (unpublished data). In each case, fewer amino acid similarities were found than in Fig. 1, and more were found the longer the segment that was used. The 35-aa translated Alu sequence alone found very few matches.

Although many years ago it was a surprise to find Alu sequences as part of the coding sequences of genes (4, 5), there are now well recognized examples (1–3, 6–9). However, in this example, Alu sequences appear to be present in a small number zinc finger genes.

## Discussion

By whatever measure is chosen, the majority of genes or amino acids are part of sets of related sequences. Seventy-three percent of the amino acids match others in regions of significant relationship, with an average of 36.5 matches for these amino acids. Eighty-three percent of the 25,193 proteins contain regions that significantly match other proteins on this list, which probably contains the great majority of human genes. Thus, the majority of genes are members of families of genes in the sense that the sequences of significant functional regions are related. The point must be made that a large fraction of the genes on this list are hypothetical or computer-derived. However, that may not much change the statistics, because the first 165 genes on the list are not hypothetical, and the fraction without significant shared domains is about the same as the total.

The cases shown in Figs. 2 and 4 show that the relationships often are restricted to local regions. For additional examples, FORTRAN is available by e-mail request. The data will be distributed on a CD for all of the 25,193 cases.

The basic issue is the mechanism that underlies these massive sets of relationships. These relationships are presumably the result of past events of regional shuffling, splicing, fusion, deletion, and duplication during evolution of protein families. During this set of processes, the individual proteins have been selected to carry out their function whether it be to produce a regulatory molecule, an enzyme, or a structural molecule. The copying, duplication, or multiplication processes have long been recognized as important (10). An overview of the resulting recognized domains is provided in ref. 11. These data show the true scale in the human genome for the 25,000 proteins recognized so far and leave no doubt that the principal process by which new gene functions arise is by making use of preexisting genes.

John Williams (California Institute of Technology) was responsible for much of the computer analysis and construction of PERL programs. Eric H. Davidson's laboratory at the California Institute of Technology supported this work.

1. Nekrutenko, A. & Li, W. H. (2001) *Trends Genet.* **17,** 619–621.
2. Lorenc, A. & Makalowski, W. (2003) *Genetics* **118,** 183–191.
3. Sorek, R. R., Ast, G. & Graur, D. (2002) *Genome Res.* **12,** 1060–1067.
4. Brownell, E., Mittereder, N. & Rice, N. R. (1989) *Oncogene* **4,** 935–942.
5. Caras, I. W., Davitz, M. A., Rhee, L., Weddell, G., Martin, D. W., Jr., & Nussenzweig, V. (1987) *Nature* **325,** 545–549.
6. Smit, A. F. (1999) *Curr. Opin. Genet. Dev.* **9,** 657–663.
7. Brosius, J. (1999) *Gene* **228,** 115–134.
8. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001) *Nature* **409,** 860–921.
9. Dagan, T., Sorek, R., Sharon, E., Ast, G. & Graur, D. (2004) *Nucleic Acids Res.* **32,** D489–D492.
10. Ohno, S. (1970) *Evolution by Gene Duplication* (Springer, Berlin).
11. Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., *et al.* (2004) *Nucleic Acids Res.* **32,** D115–D119.