15. C. Navarrete, *The Olmec Rock Carvings at Pijijiapan, Chiapas, Mexico and Other Olmec Pieces from Chiapas and Guatemala* (Paper 35, New World Archaeological Foundation, Provo, UT, 1974).
16. M. D. Coe, in *Origins of Religious Art and Iconography in Preclassic Mesoamerica*, H. B. Nicholson, Ed. (Univ. of California Press, Berkeley, CA, 1976), pp. 107–122; J. Marcus, *Annu. Rev. Anthropol.* **5**, 35 (1976); M. Ayala, in *Antropología e Historia de los Mixe-Zoques y Mayas (Homenaje a Frans Blom)*, L. Ochoa and T. A. Lee, Jr., Eds. (Universidad Nacional Autónoma de México, Coyoacán, Mexico, 1983), pp. 175–221; J. S. Justeson, W. M. Norman, L. Campbell, T. Kaufman, *The Foreign Impact on Lowland Mayan Language and Script* (Publ. 53, Middle American Research Institute, New Orleans, LA, 1985); S. Meluzin, in *The Periphery of the Southeastern Maya Realm*, G. Pahl, Ed. (UCLA Latin American Center, Los Angeles, 1987), pp. 67–112.
17. M. Ayala, in (*16*), p. 197; S. Meluzin, in (*16*), p. 69; J. Justeson, report in a University of California at Berkeley seminar on non-Mayan scripts of Mesoamerica (1970), directed by J. A. Graham.
18. G. W. Lowe, in *The Origins of Maya Civilization*, R. E. W. Adams, Ed. (Univ. of New Mexico Press, Albuquerque, NM, 1977), pp. 197–248; J. S. Justeson, *World Archaeol.* **17**, 437 (1986); _____ and P. Mathews, *Visbl. Lang.* **24**, 88 (1990); B. Stross, *ibid.*, p. 38. However, the text is instead addressed in terms of Mayan vocabulary and grammar by M. D. Coe [as cited in (*16*)] and by L. B. Anderson [*The Writing System of La Mojarra and Associated Monuments* (privately printed, Washington, DC, 1991)].
19. M. Macri, in *Literacies: Writing Systems and Literate Practices*, D. L. Schmidt and J. S. Smith, Eds. (Davis Working Papers in Linguistics, vol. 4, Department of Linguistics, University of California at Davis, CA, 1991), pp. 11–23.
20. L. Campbell and T. Kaufman, *Am. Antiq.* **41**, 80 (1976).
21. Based on part of this evidence, the *wu* reading was proposed by T. Kaufman in a working group on La Mojarra Stela 1 at the meeting of the American Anthropological Association, Phoenix, AZ, 16 to 20 November 1988.
22. Ergative is the grammatical category of the agent of a transitive verb and, in Mixe-Zoquean languages, the possessor of a noun.
23. B. Stross, in (*18*), pp. 48–51.
24. The basis for our calendrical framework was presented by J. S. Justeson at the Workshop on La Mojarra Stela 1, University of California at Santa Barbara, CA, April 1989.
25. J. S. Justeson and P. Mathews, in (*18*), pp. 97 and 99; also presented by J. S. Justeson at the workshop in (*24*). The 'bloodletting' meaning of MS132 was identified by Mathews.
26. J. S. Justeson, W. M. Norman, L. Campbell, T. Kaufman, in (*16*), pp. 38–44; L. B. Anderson, in (*18*). For a contrary view, see J. E. S. Thompson, *Archaeology of Southern Mesoamerica: Part 2*, vol. 3 of *Handbook of Middle American Indians* (Univ. of Texas Press, Austin, 1965), p. 651.
27. J. S. Justeson and P. Mathews, in (*18*), p. 114.
28. A Mixe-Zoquean source for this pair of values in Mayan had been postulated by J. S. Justeson before the epi-Olmec sign was known [J. S. Justeson, W. M. Norman, L. Campbell, T. Kaufman, in (*16*), p. 44].
29. J. S. Justeson and T. Kaufman, *The Decipherment of Epi-Olmec Hieroglyphic Writing and Mixe-Zoquean Comparative Linguistics* (Univ. of Oklahoma Press, Norman, OK, in press).
30. F. Winfield Capitaine, in (*12*), p. 16.
31. All who have studied La Mojarra Stela 1 owe a tremendous debt to L. Wagner, G. Stuart, and F. Winfield Capitaine, who were instrumental in effecting its unrestricted dissemination. G. Stuart also produced the drawings of the text and monument that have been the indispensable basis for all subsequent work and provided us with access to his unpublished photographs. We began our joint work on epi-Olmec writing in March 1991 in the context of a workshop organized by M. Macri under the auspices of the University of Texas Workshop on Maya Hieroglyphic Writing. Travel support for our collaboration has been provided in part by the Natural Language Group at IBM Research (J.S.J.) and the Texas workshop (T.K.). We thank the National Geographic Society for funding the continuation of this research, in particular fieldwork on Mixe-Zoquean languages.

Ancient Conserved Regions in New Gene Sequences and the Protein Databases

Philip Green,* David Lipman, LaDeana Hillier, Robert Waterston, David States, Jean-Michel Claverie

Sets of new gene sequences from human, nematode, and yeast were compared with each other and with a set of *Escherichia coli* genes in order to detect ancient evolutionarily conserved regions (ACRs) in the encoded proteins. Nearly all of the ACRs so identified were found to be homologous to sequences in the protein databases. This suggests that currently known proteins may already include representatives of most ACRs and that new sequences not similar to any database sequence are unlikely to contain ACRs. Preliminary analyses indicate that moderately expressed genes may be more likely to contain ACRs than rarely expressed genes. It is estimated that there are fewer than 900 ACRs in all.

Understanding the functions and structures of the array of proteins expressed in living organisms is a fundamental goal of molecular biology. Our hope of attaining this goal stems largely from the unifying theme of shared evolutionary ancestry: related organisms have similar proteins and, within an organism, different proteins of related function are often wholly or partly similar in sequence, reflecting gene duplication and exon shuffling (*1*) during evolution. Such similarities can provide important functional insights, and consequently an important step in characterizing any newly sequenced gene is to compare its encoded protein sequence with the protein sequence databases in order to look for conserved regions shared with known proteins.

The present study uses extensive new sets of gene sequences to address several general questions about conserved regions: how many of these regions exist, what fraction has been discovered, and what proportion and types of proteins contain them. We focus on ancient conserved regions, or ACRs, detected through similarities between proteins from distantly related organisms. Over long evolutionary periods the less constrained portions of the sequences will have significantly diverged; consequently, the regions of similarity are usually those of greatest structural or functional significance. ACRs often correspond to specific domains (or motifs) present in a variety of proteins, such as zinc finger DNA binding domains (*2*), or to enzyme active sites, but they can also comprise most or all of the sequence of a single highly conserved protein or protein family, such as actins and histones. Conserved regions of all of these types have been extensively cataloged (*3*, *4*). Because the degree of similarity between two related proteins reflects not only the amount of time since their last common ancestor but also their rates of sequence evolution, which can vary greatly for different proteins (*5*), not all proteins need contain ACRs.

The precise definition of an ACR depends on its required age and distribution among organisms and on the method used to detect sequence similarities. The present study involves ACRs that antedate the radiation of the major animal phyla [some 580 to 540 million years ago (*6*)] and that are present in diverse eukaryotes. We detected similarities by using the sequence alignment program BLAST (*7*) with a score cutoff sufficiently high to distinguish confidently true homologies from background in database searches (*8*). Figure 1 shows a representative BLAST alignment at this score level. Typically, a BLAST comparison of two related proteins reveals several (gap-free) aligned segments, separated by unaligned regions; in such cases we considered the entire collection of aligned segments to constitute a single conserved region, provided the segments always tended

P. Green, L. Hillier, and R. Waterston are in the Genetics Department, Washington University Medical School, St. Louis, MO 63110. D. Lipman, D. States, and J.-M. Claverie are at the National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894.

*To whom correspondence should be addressed.

to be conserved together in homologies with other sequences.

We analyzed four extensive collections of unselected gene sequences that have recently become available: partial cDNA sequences [expressed sequence tags (ESTs)] from human brain (9) and the nematode *Caenorhabditis elegans* (10) and gene sequences inferred from tracts of genomic DNA of *C. elegans* (11) and the yeast *Saccharomyces cerevisiae* (12). To investigate conserved regions that antedate the prokaryote-eukaryote divergence, we also analyzed *Escherichia coli* gene sequences from the databases; these now comprise about 40% of the *E. coli* genome and so must constitute a reasonably representative sample from this organism.

Fig. 1. BLAST alignment detecting an ACR from the present study [Src homology 3 domain (*33*)]. The query and subject sequences are encoded by a predicted gene in *C. elegans* genomic cosmid clone B0303 and by the ABP1 gene on yeast chromosome 3, respectively. Abbreviations for the amino acid residues are as follows: A, Ala; D, Asp; E, Glu; F, Phe; G, Gly; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.

```
Score = 75 (36.9 bits), Expect = 7.8 10-5, P = 7.8 10-5
Identities = 14/37 (37%), Positives = 21/37 (56%)

Query   553 ATAIYDYNSNEAGDLNFAVGSQIMVTARVNEEWLEGE 589
            ATA YDY++ E  +L F     I+    V+++W  GE
Sbjct   537 ATAEYDYDAAEDNELTFVENDKIINIEFVDDDWWLGE 573
```

ACRs in the Sequence Sets

The new sequence sets have previously been searched against the protein databases to identify homologies (*9–12*); however, such homologies do not necessarily represent ACRs unless they involve sequences from organisms in different phyla. Table 1 indicates the numbers of ACRs in each set revealed by cross-phylum homologies against the SWISS-PROT database (*13*). (We will refer to ACRs represented in current database protein sequences as database ACRs.) The estimates of the fraction of coding sequences with database ACRs in Table 1 are crude, particularly for the human ESTs, because it is difficult to estimate the number of coding sequences (*14*). Furthermore, the ESTs usually do not contain complete coding units, with the result that the full gene may contain a conserved region when the EST does not. The results are nevertheless roughly consistent for the various sets and suggest that ~20 to 40% of the coding sequences in these sets contain database ACRs. Use of less conservative matching criteria would increase this percentage somewhat, but at the expense of a higher rate of false positives (*8*).

We then asked whether we could find additional conserved regions by comparing the sequence sets with each other. Any match between sequences from two of these organisms represents an ACR, by definition, and the ACRs found in this way should constitute a random selection of those present in the sets. Table 2 shows the results of these pairwise comparisons. Because over half of the sequences in each set had no database match, we expected that more than half of the matches between sets might represent new ACRs. Surprisingly, very few new ACRs were detected. For example, only 42% of the C. *elegans* ESTs match a database sequence, yet over 90% of the C. *elegans* ESTs matching human ESTs, yeast open reading frames (ORFs), or *E. coli* genes also match a database sequence. We detected a total of 54 distinct ACRs by comparing the new sequence sets with each other, of which 49 (91%) were already represented in known proteins (Table 3).

The *E. coli* comparisons show an even higher proportion of database ACRs. This suggests that prokaryote-eukaryote ACRs,

Table 1. ACRs detected by comparison of the sequence sets to the protein sequence database. Each set was searched against SWISS-PROT (*13*) to identify all matches [as defined by the criteria (*8*)] with sequences from a different phylum (for example, for the human ESTs only matches against nonchordate sequences were counted). The numbers of ACRs are what remain after redundancies have been eliminated; this was done in semi-automated fashion by performing all pairwise comparisons of matched sequences and clustering into groups those sequences that share similar subsequences. For this purpose, database sequences known on biological grounds to be homologous were clustered even if the matching criteria (*8*) were not strictly met.

Set	*n*	Coding sequences	Sequences with ACRs*	ACRs
Human ESTs†	2644	600 to 1200	197 (16 to 33%)	103
C. elegans ESTs†	1472	1370	570 (42%)	240
C. elegans genes‡	234	234	74 (32%)	59
Yeast ORFs§	182	182	43 (24%)	35
E. coli genes‖	1916	1916	439 (23%)	266

*Values in parentheses indicate the estimated percentage of coding sequences with ACRs. †The number of coding sequences in each EST set was estimated as described (*14*). For the comparisons, human EST sequences were conceptually translated in all six possible reading frames and worm ESTs in the three top-strand frames, eliminating ORFs shorter than 30 codons. ‡Predicted from the *C. elegans* genomic sequence using the program Genefinder [(*11*); P. Green and L. Hillier, unpublished results]. §The 182 probable coding sequences (*12*) are the ORFs having an in-frame ATG followed by at least 300 nucleotides. ‖Translated *E. coli* sequences were taken from SWISS-PROT. Values shown reflect matches with eukaryotic sequences. If matches with sequences from Gram-positive eubacteria and archaebacteria are also counted, 728 (38%) of *E. coli* genes have ACRs, and there are 396 distinct ACRs. Many of the additional ACRs may be prokaryote-specific, however.

Table 2. ACRs detected by comparison of the sequence sets to each other. Shown for each comparison are the number of sequences in each set that match some sequence in the other set, the number of distinct ACRs detected by these matches (counting homologs only once), the number of these ACRs that are homologous to a known sequence in the National Center for Biotechnology Information nonredundant protein database, and the fraction of distinct ACRs that are present in the database, respectively. The criteria for matching and for clustering homologous ACRs into groups were as described [(*8*) and Table 1]. For the *E. coli* comparisons, database ACRs are those homologous to a database sequence from a Gram-positive bacterium, archaebacterium, or eukaryote.

Sets compared	Matching sequences	ACRs	ACRs in database
C. elegans ESTs, human ESTs	77, 66	34	31 (91%)
C. elegans ESTs, yeast ORFs	23, 13	9	8 (89%)
C. elegans genes, human ESTs	17, 17	12	12 (100%)
C. elegans genes, yeast ORFs	6, 4	4	3 (75%)
Human ESTs, yeast ORFs	14, 13	10	10 (100%)
Total* (new set comparisons)		54	49 (91%)
Total* (eukaryote-specific ACRs)†		29	24 (83%)
C. elegans ESTs, *E. coli* genes	136, 100	57	56 (98%)
C. elegans genes, *E. coli* genes	18, 23	16	15 (94%)
Human ESTs, *E. coli* genes	21, 33	17	17 (100%)
Yeast ORFs, *E. coli* genes	17, 46	16	16 (100%)
Total* (*E. coli* comparisons)		81	79 (98%)

*After removal of redundant ACRs. †Omits all ACRs homologous to any prokaryotic sequence in database (see text).

which are found in both prokaryotes and eukaryotes and therefore antedate their divergence [perhaps 3.5 billion years ago (6)], should be considered separately from eukaryote-specific ACRs, which on current evidence (15) are present only in eukaryotic proteins and so are probably of more recent origin (for example, the conserved domains of cytoskeletal proteins). The *E. coli* comparisons specifically detect prokaryote-eukaryote ACRs and indicate that 98% of these may be present in the database. Consequently, the five ACRs not represented in the database that were found by comparing the new sequence sets are more likely to be eukaryote-specific. This assumption then leads to an estimate that 83% of the eukaryotic-specific ACRs in these sets are present in the databases (Table 2).

Table 3. Database ACRs detected by comparison of new sequence sets. Where possible, conserved region designations follow BLOCKS (*4*) and PROSITE (*3*). H, human ESTs; WE, worm ESTs; WG, worm genomic; Y, yeast ORFs. ACRs from *E. coli* comparisons are not included here.

ACR	Comparison	Score range
Eukaryote-specific ACRs		
Actins	WE, H	87 to 529
Adenylate cyclases*	WG, H	80
β-Transducin family	WE, H, Y	77 to 215
DIFF6 protein (mouse)*	WE, H, Y	94 to 135
EF-hand calcium-binding domain	WG, H; WE, H	76 to 225
Epidermal growth factor–like domain cysteine pattern	WG, H	96 to 99
Epoxide hydrolase*	WE, H	80
Eukaryotic RNA-binding region RNP-1	WE, H	123
Fructose bisphosphate aldolase class I	WE, H	217
Guanosine diphosphate dissociation inhibitor for SMG P25A*	WE, H	111
Gelsolin*	WE, H	85
G10 protein (*Xenopus laevis*)*	H, Y	120
Hexokinases	H, Y	89 to 155
Intermediate filament proteins	WE, H	88 to 114
Kinesins	WG, H	151 to 237
Neurotransmitter transporters	WG, H	101
Phorbol esters-diacylglycerol binding domain	WG, H	95 to 100
Protein kinase catalytic domain	WE, H, Y	76 to 199
Ras-like guanosine triphosphatase family*	All	75 to 233
SEC7 homolog*	WE, H	76
Src-homology 3 domain*	WG, H; WG, Y	75 to 108
Talin*	WG, H	104
Tubulins	WE, H	75 to 429
Ubiquitin	WE, H	146 to 336
Zinc finger, C3HC4 type	WG, Y	79 to 113
Prokaryote-eukaryote ACRs		
3-Hydroxyacyl–coenzyme A dehydrogenase	WE, H; WG, H	98 to 99
Adenosine triphosphate–binding proteins active transport family	H, Y; WE, Y	75 to 107
Aminoacyl-transfer RNA synthetase class II	WE, Y	84
Cytidine diphosphate–diacylglycerol-serine *O*-phosphatidyltransferase*	H, Y	78
Citrate synthase	WE, Y	207 to 372
Cyclophilin-type peptidyl-prolyl *cis-trans* isomerase	H, Y	106
Cytochrome c oxidase subunit I, copper B binding region	WE, H	218
E1-E2 adenosine triphosphatases phosphorylation site	WE, H	84 to 366
Enolase	WE, H	331
Glyceraldehyde 3-phosphate dehydrogenase	WE, H	223 to 234
Guanosine triphosphate–binding elongation factors	WE, H	80 to 602
HSP70 family	WE, H; WG, H	89 to 246
HSP90 family	WE, H	217
Insulinase family	WE, H	78 to 195
Malate–L-lactate dehydrogenases	WE, H	161 to 253
NifH-FraXC family	WG, H	75
Phosphoglycerate kinase	H, Y	132
Pyruvate dehydrogenase E1 (α subunit)*	WE, H	124 to 152
Pyruvate kinase	WE, H	162
Ribosomal protein L3	WE, H	172
Ribosomal protein P0 (L10E)*	WE, H	250 to 257
Serine proteases, subtilase family	WE, H	85
Thioredoxin family	WE, Y	76 to 170
Triosephosphate isomerase	WE, H	189

*Denotes ACR not present in BLOCKS.

Possible biases. If the ACRs identified by these comparisons were a random sample of all ACRs, then we could conclude that roughly 98% of all prokaryote-eukaryote ACRs and 83% of eukaryote-specific ACRs are already represented in the databases. For two reasons, however, the sample is not random, and consequently these values are likely to be overestimates. First, ACRs that are present in many different proteins are more likely to be represented in these sets and are also more likely to be present in the databases. This creates a bias that we have only partly compensated for by eliminating redundancies. However, inspection of the lists of database ACRs (Table 3) suggests that this bias is unlikely to be large: although several ACRs, such as the ras guanosine triphosphatase domain and the protein kinase catalytic domain, are found in extensive protein families, the majority are not currently believed to be.

A second bias is that the EST sets, like the sequence databases, tend to favor moderately expressed genes over rarely expressed ones. This does not affect our conclusions regarding prokaryote-eukaryote ACRs, which are supported by comparisons of genomic sets (16). It could affect our estimates for the eukaryote-specific ACRs (most of which were found in comparisons involving the EST sets), but even here we believe that the bias is not large. First, although the numbers are small, the comparison of the (unbiased) yeast and *C. elegans* genomic sets (Table 2) is consistent with the estimate for all eukaryote-specific ACRs. Second, the list of eukaryote-specific ACRs (Table 3) is hardly dominated by ACRs in highly expressed proteins (although it does include some).

Third, the cDNA libraries used here should be somewhat less biased toward highly expressed genes than are many other libraries (17), and in fact the compositions of the EST sets appear quite diverse [as is evident from the lists of homologies identified in the original studies (9, 10)]. The *C. elegans* ESTs, for example, are thought to represent 1,194 distinct genes, or about 8% of the estimated 15,000 genes in this organism (10, 11). The number of unique database ACRs per *C. elegans* EST is about 0.16, or 0.20 if overlapping ESTs are counted once (Table 1); this compares favorably with the value of 0.25 for the *C. elegans* genes inferred from genomic sequence, particularly in view of the much larger size of the EST set (18). Moreover, the number of distinct database ACRs in the *C. elegans* ESTs is about one-third of the

total estimated number of ACRs in SWISS-PROT (see below). Furthermore, it is clear that the EST sets are sampling from a reasonably large set of genes because many ESTs do not have homologs in the database.

Fourth, preliminary analyses (below) suggest that rarely expressed genes may be less likely to contain ACRs. This would tend to diminish the possible bias in the estimates caused by underrepresentation of such genes in the EST sets.

It therefore appears likely that most ACRs are represented in the current protein databases. This conclusion is supported by the finding that the rate of discovery of new ACRs through cross-phylum sequence matches has declined dramatically over the past few years, in both relative and absolute terms (*19*), despite an exponential increase in sequence data. It is perhaps not surprising that most prokaryote-eukaryote ACRs would have been discovered because the last common ancestor of prokaryotes and eukaryotes may have been a relatively simple organism (*20*) and because this subset of ACRs must have been more highly conserved to retain detectable homology across several billion years. It is surprising that most eukaryote-specific ACRs should have been found; presumably, this reflects the relatively small number of ACRs (see below) and the enormous increase in sequence data over the last decade.

Sensitivity of analyses. Our results may depend to a certain extent on the sensitivity of the method used to detect ACRs. We looked for ACRs that are highly enough conserved to be detectable by standard database searches using a pairwise sequence alignment program. Although this criterion is the one most commonly used to discover homologies, there are other more sensitive methods involving comparisons of multiple sequences (*4, 21*), which may have the potential to reveal additional, somewhat less highly conserved regions. The use of such methods would certainly increase the proportion of sequences found to contain database ACRs. Additional matches between the new sequence sets would presumably be detected, as well as additional corresponding database matches, making it difficult to predict how the proportion of ACRs that are represented in known proteins would change, if at all. It is possible that this proportion would decrease (*22*); however, we found that the use of a simpler method to increase sensitivity [the Smith-Waterman algorithm (*23*)] actually increased the proportion slightly when the *E. coli* and yeast sets were compared: a single additional ACR was found and was also present in known proteins. In any case, it remains to be shown in practice how many additional ACRs the multiple sequence comparison methods can actually detect.

Table 4. Homologies and ORF lengths of singly and multiply represented *C. elegans* ESTs. [Doubly represented ESTs overlap one other EST, and triply represented ESTs at least two others, by the criteria described in (*10*).] ESTs were subgrouped into those with database ACRs and those without database ACRs (values given in parentheses), and within each subgroup matches with the other sets were counted. Homology criteria were as described [(*8*) and Table 1].

	EST representation		
	Single	Double	Triple
n	310 (675)	144 (154)	116 (73)
Similar to human EST	34 (1)	23 (2)	17 (0)
Similar to yeast ORF	15 (1)	5 (0)	2 (0)
Similar to *E. coli* gene	67 (1)	46 (0)	22 (0)
Similar to another* *C. elegans* EST	92 (41)	59 (13)	46 (24)
Having ORFs >300 nucleotides	259 (494)	115 (105)	92 (54)

*Nonoverlapping.

How Many ACRs?

Conserved protein regions have been extensively cataloged in PROSITE (*3*) and BLOCKS (*4*). Of 91 distinct database ACRs in the *C. elegans* and *S. cerevisiae* genomic sequences, 60 correspond to known conserved regions in version 5.0 of BLOCKS (*24*). Thus, we estimate that roughly two-thirds of all database ACRs are represented in BLOCKS. There are a total of 559 conserved regions in BLOCKS (*24*), of which 78 currently appear to be restricted to prokaryotic proteins and so do not correspond to ACRs by our definition. If all of the remaining 481 conserved regions represent ACRs (*25*), then the total number of ACRs currently in SWISS-PROT should be about 481/0.66 = 730. About 40% of these, or 300, should be prokaryote-eukaryote ACRs (*26*).

The assumption that 85% of ACRs are represented in the database then yields an estimate of roughly 860 (that is, 730/0.85) for the total number of ACRs. Of these, roughly 290 either are not yet represented in the database or are represented only in sequences from a single phylum (*27*), leaving an estimated 570 that are currently present in multiple phyla in the database. The latter figure is consistent with the results of recent studies (*19, 28*) that compared sequences in SWISS-PROT with each other to detect cross-phylum matches and found 500 to 600 different ACRs (using slightly different criteria).

These numbers are somewhat dependent on the method used to detect ACRs and therefore are only approximate. Nonetheless, we can begin to generate a preliminary picture of the distribution of ACRs in the genome. *C. elegans*, for example, is estimated to have about 15,000 genes (*10, 11*). Because about a third of these have database ACRs (Table 1), on average a given database ACR appears in about seven genes (5000/730) in this organism (this calculation ignores genes with multiple ACRs, which appear to be relatively infrequent). There is likely considerable variation among ACRs, with some represented only once and others represented many times; a more detailed picture will emerge as the sequencing projects progress. It will also be of interest to learn what proportion of the ACRs are specific to metazoans.

Expression Level and Degree of Conservation

To better understand the impact of expression level bias in the EST sets, we looked for a possible relation between expression level and ACR presence. Because detailed expression data on these clones are not yet available, we assumed that to a first approximation genes represented in multiple independent clones in the cDNA libraries are, on average, expressed at higher levels than singly represented genes. Analyses were confined to the *C. elegans* ESTs (*29*), which were classified as singly represented (not overlapping any other EST) or multiply represented (overlapping at least one other EST). We found (Table 4) that database ACRs are present in a substantially higher fraction of the multiply represented ESTs (260/487, or 53%) than of the singly represented ESTs (310/985, or 31%). A similar trend holds for the *C. elegans* ACRs detected by similarity to the other sequence sets (*30*). Moreover, multiply represented ESTs have generally higher similarity scores with their distant homologs in the database than do singly represented ESTs (Fig. 2). The higher proportion of ACRs among multiply represented ESTs thus appears to be at least in part a consequence of their generally stronger similarities with distantly related genes and cannot simply be explained by a bias in the database itself toward moderately to highly expressed genes (*31*).

These results suggest that moderately expressed proteins have, on average, been more highly conserved in sequence over long evolutionary periods than have rarely expressed ones and in particular are more likely to contain ACRs. This is presumably

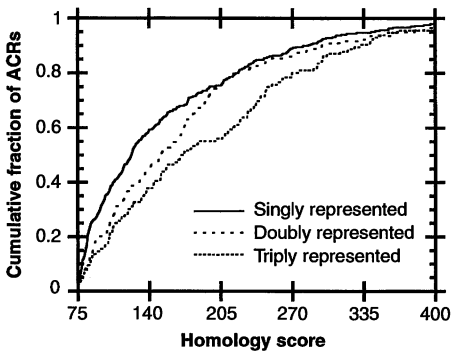

Fig. 2. Distribution of homology scores for database ACRs in singly and multiply represented *C. elegans* ESTs. For each EST having a cross-phylum match against SWISS-PROT, the average score of all such matches was taken to indicate the degree of conservation of the corresponding ACR. The cumulative fraction of ACRs having average scores less than a given value is plotted. Relatively more of the multiply represented ESTs have average scores exceeding any given value.

attributable in part to higher selective pressures to optimize the activities and structures of these proteins and to minimize undesired interactions with other cellular components. Given the indirectness of our method of assessing expression level, more detailed expression data on these clones will be required to confirm and accurately quantify this correlation.

Sequences Without ACRs

An early finding of the genome sequencing projects was that the majority of genes are not similar to anything in the databases (*11*, *12*). It has usually been assumed that this reflects the relative incompleteness of the databases rather than the absence of highly conserved regions in these genes. This assumption now appears incorrect. Because 30% or fewer of the genes in the genomic sets we analyzed contain database ACRs, and perhaps 85% of ACRs are present in the databases, the fraction of genes that contain ACRs is roughly 40% (0.30/0.85) or less. The other 60%—or over 90% of those sequences that are not currently similar to a distantly related sequence in the databases—do not have ACRs and must therefore correspond to proteins or protein regions that either evolved more recently than the metazoan radiation or evolved prior to it but have not been strongly conserved (*5*). In either case, they are unlikely to have strong similarities to any genes from distantly related organisms. For these sequences, homologies will be detectable only with the use of more sensitive methods of analysis or by comparisons with genes from more closely related organisms.

Many of these genes may have ancient functions despite their lack of sequence conservation. It is unlikely that the sequence requirements for a minimally active protein of any given function could be particularly stringent; otherwise, given the improbability of a specific sequence of any significant length arising solely by chance mutation, an appropriate substrate for selection to begin acting upon would never have arisen. Although optimization of activity can entail much more stringent sequence requirements, such optimization may only have been strongly selected for in a minority of the proteins in an organism. Thus, the majority of protein sequences may be relatively unconstrained and as a result may be drifting too rapidly to retain detectable similarities over long evolutionary periods. For this reason, one should not assume that ACRs necessarily represent all of the ancestral functional domains. Nor do they correspond to the universe of ancestral exons (*32*) because the majority of exons do not appear to be highly conserved. In fact, the differential rate of evolution of different protein regions considerably complicates the task of estimating the ancestral exon number.

In summary, it appears that the number of ACRs is relatively small—far smaller than the number of genes in a eukaryotic genome—and that most ACRs are represented among currently known proteins. We would emphasize, however, that more sequence data will be required to improve our understanding of conserved protein regions. The estimates above suggest that roughly one-third of ACRs have not yet been discovered because they are represented in only one phylum (or not at all) in the current databases. Detection of less highly conserved ACRs may only be possible when they are represented in multiple distantly related sequences. Finally, to increase our understanding of sequences that lack ACRs, it will be important to acquire sequence information from closely related organisms.

REFERENCES AND NOTES

1. W. A. Gilbert, *Science* **228**, 823 (1985).
2. J. M. Berg, *ibid.* **232**, 485 (1986).
3. A. Bairoch, *Nucleic Acids Res.* **20**, 2013 (1992).
4. S. Henikoff and J. G. Henikoff, *ibid.* **19**, 6565 (1991).
5. R. F. Doolittle, *Protein Sci.* **1**, 191 (1992).
6. A. H. Knoll, *Science* **256**, 622 (1992).
7. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. Lipman, *J. Mol. Biol.* **215**, 403 (1990). BLAST has been shown [W. R. Pearson, *Genomics* **11**, 635 (1991); S. F. Altschul *et al.*, *ibid.*, p. 408] to have comparable sensitivity to two other commonly used sequence alignment methods, FASTA [W. R. Pearson and D. J. Lipman, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 2444 (1988)] and the Smith-Waterman dynamic programming method (*23*). In the present study, all BLAST comparisons were done with the use of conceptual translations of the DNA sequences.
8. Matches with scores of at least 75 obtained with the PAM120 matrix [M. O. Dayhoff, R. M. Schwartz, B. C. Orcutt, in *Atlas of Protein Sequence and Structure* (National Biomedical Research Foundation, Washington, DC, 1979), vol. 5, suppl. 3, pp. 345–352; S. F. Altschul, *J. Mol. Biol.* **219**, 555 (1991)] were regarded as significant. For an average protein of length 250 amino acid residues searched against a database of 20 million residues, a PAM120 score of 75 corresponds to a *P* value of 0.006. Two sequences sharing regions of highly biased amino acid composition (such as charged residue domains) can have a high similarity score owing to convergence and not homology. We prefiltered all sequences with a program that eliminates most such regions (XNU) (J.-M. Claverie and D. States, *Comput. Chem.*, in press). Use of such programs may occasionally mask homologies but helps avoid false positives. With these criteria very few of the matches are likely to be spurious, although some biologically significant matches may be missed.
9. M. D. Adams *et al.*, *Science* **252**, 1651 (1991); M. D. Adams *et al.*, *Nature* **355**, 632 (1992).
10. R. Waterston *et al.*, *Nat. Genet.* 1, 79 (1992).
11. J. Sulston *et al.*, *Nature* **356**, 37 (1992); *C. elegans* sequencing consortium, unpublished data. Assembled genomic sequence comprising 1,085,730 nucleotides (in 38 cosmids) from chromosome 3 was used in the present study.
12. S. G. Oliver *et al.*, *ibid.* **357**, 38 (1992).
13. A. Bairoch and B. Boeckmann, *Nucleic Acids Res.* **19**, 2247 (1991). We used version 23 and excluded all sequences from organisms of unknown phylogenetic classification (for example, viruses).
14. The EST sets clearly contain some ESTs with insufficient accurate coding sequence to detect homologies, including contaminants, repetitive element transcripts, and ESTs with sequencing errors in critical regions or with a significant fraction of untranslated sequence. We estimate that only about 7% of the *C. elegans* ESTs are noncoding in the above sense: about 82% (466/570) of ESTs with homologies to known proteins have a top-strand ORF greater than 300 nucleotides in length, as compared with 72% (653/902) of ESTs without such homologies. If we assume that the noncoding ESTs do not have ORFs of this length, then they should represent about [1 − (0.72/0.82)] = 12% of the ESTs without homologies, or about 7% of the total set of ESTs. Two other lines of evidence also suggest that there are relatively few noncoding *C. elegans* ESTs: (i) virtually all the clones hybridize to one or a small number of locations in the *C. elegans* physical map [(*10*); A. Coulson, R. Shownkeen, C. Huynh, R. Waterston, unpublished results], which implies that there are few if any exogenous DNA or repetitive DNA clones, and (ii) the homology rates for ESTs and for genes inferred from genomic sequence are roughly consistent (Table 1). It has previously been estimated (*9*) that from 50 to 85% of the human ESTs without homologies to known proteins are noncoding.
15. Some eukaryote-specific ACRs may be present in prokaryotes but may not yet have been detected as a result of their greater divergence.
16. We regard the *E. coli* set as effectively genomic because it represents 40% of the genome.
17. The human brain has greater transcriptional complexity than other human tissues [J. G. Sutcliffe, *Annu. Rev. Neurosci.* **11**, 157 (1988)]; the *C. elegans* library was partially normalized by repeated negative screening with probes consisting of total labeled cDNA and of pools of clones already selected from the library (*10*).
18. Because the universe of ACRs must be finite, as one enlarges the set of sequences, the redundancy of ACRs detected must increase, and therefore the number of unique ACRs per sequence must decrease.
19. D. States, L. Hunter, N. Harris, D. Lipman, in preparation.
20. C. R. Woese, in *Archaebacteria*, O. Kandler, Ed. (Fischer, Stuttgart, Germany, 1982), pp. 1–17.
21. M. Gribskov, A. D. McLachlan, D. Eisenberg, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 4355 (1987); J. Posfai, A. S. Bhagwat, R. J. Roberts, *Gene* **74**, 261 (1988); S. F. Altschul and D. J. Lipman, *Proc. Natl. Acad. Sci. U.S.A.* **87**, 5509 (1990); G. J. Barton and M. J. Sternberg, *J. Mol. Biol.* **212**, 389 (1990);

A. Danckaert, C. Chappey, S. Hazout, *Comput. Appl. Biosci.* **7**, 509 (1991).
22. Given the apparent relation between expression level and degree of conservation, it might be expected that rarely expressed genes are relatively more likely to have weakly conserved, marginally detectable ACRs. Such genes and ACRs may be underrepresented in the current databases.
23. T. Smith and M. S. Waterman, *J. Mol. Biol.* **147**, 195 (1981).
24. In BLOCKS, a conserved region (in our sense) is represented as one or more "blocks," each consisting of a set of (gap-free) aligned segments from homologous proteins in SWISS-PROT. We considered a database ACR to correspond to a known conserved region in BLOCKS if (i) at least one of the BLAST alignments that detects the ACR involves a SWISS-PROT sequence included in a block and (ii) a BLAST-aligned segment from that sequence overlaps a segment in the block.
25. Some conserved regions in BLOCKS may not meet our criteria for an ACR because they are restricted to a single eukaryotic phylum. This would cause our estimate for the number of database ACRs to be too high; however, the close agreements (see following) of the estimated number of prokaryote-eukaryote ACRs with results for the *E. coli* set (Table 1) and of the estimated number of database ACRs present in multiple phyla with the results of direct database comparisons suggest that the discrepancy cannot be large.
26. This estimate is based on the observation that 37 of the 91 yeast and *C. elegans* genomic database ACRs currently match prokaryote sequences in SWISS-PROT.
27. Twenty of the yeast and *C. elegans* genomic ACRs match sequences from only a single phylum in the database and do not match any conserved region in BLOCKS. This suggests that roughly 730(20/91) = 160 of database ACRs are represented in only a single phylum. The number of ACRs with no representative in the database is roughly (860 − 730) = 130.
28. J.-M. Claverie, in preparation.
29. All clones in the *C. elegans* library were sequenced from the 5′ end and were systematically size selected, thus increasing the likelihood that different clones deriving from the same gene would be detectable by sequence overlap. A significant fraction of these cDNAs are essentially full length because about 14% of the ESTs include part of a transspliced leader sequence at the 5′ end. (It is unknown what fraction of *C. elegans* transcripts are transspliced).
30. The homologies between nonoverlapping *C. elegans* ESTs represent regions that have been conserved (over unknown evolutionary periods) within *C. elegans*. The distribution of these regions among ESTs with and without database ACRs suggests that a high proportion (but not all) of conserved protein regions that are multiply represented within *C. elegans* are actually ACRs.
31. The comparisons between the singly represented ESTs and the other new sequence sets (Table 4) indicate that over 90% of ACRs within the singly represented EST set are present in the database; thus, singly represented ESTs have relatively fewer ACRs rather than only a smaller proportion of database ACRs. In particular, the difference in rates cannot be attributed to underrepresentation of rarely expressed genes in the databases. Nor can it be attributed to a higher fraction of noncoding ESTs in the singly represented set because the ORF length data in the two sets (Table 4) suggest that the proportion of noncoding ESTs among ESTs without ACRs is about the same (12%) for the singly represented ESTs as for the multiply represented ESTs (*14*). Moreover, neither of these alternative hypotheses could account for the difference in homology score distributions of Fig. 2, which is based solely on the database ACRs.
32. R. L. Dorit, L. Schoenbach, W. Gilbert, *Science* **250**, 1377 (1990).
33. M. L. Stahl, C. R. Ferenz, K. L. Kelleher, R. W. Kriz, J. L. Knopf, *Nature* **332**, 269 (1988).
34. Partly supported by grants from the National Center for Human Genome Research. We thank R. Durbin, M. Olson, T. Schedl, and D. Schlessinger for helpful comments on the manuscript. This paper was submitted for review on 21 May 1992 and accepted for publication on 9 December 1992.