Gene Families: The Taxonomy of Protein Paralogs and Chimeras

Steven Henikoff,* Elizabeth A. Greene, Shmuel Pietrokovski, Peer Bork, Teresa K. Attwood, Leroy Hood

Ancient duplications and rearrangements of protein-coding segments have resulted in complex gene family relationships. Duplications can be tandem or dispersed and can involve entire coding regions or modules that correspond to folded protein domains. As a result, gene products may acquire new specificities, altered recognition properties, or modified functions. Extreme proliferation of some families within an organism, perhaps at the expense of other families, may correspond to functional innovations during evolution. The underlying processes are still at work, and the large fraction of human and other genomes consisting of transposable elements may be a manifestation of the evolutionary benefits of genomic flexibility.

Linnaeus introduced a universal classification system of living things that was able to organize the enormous complexity of biological relationships. A universal gene classification system presents a similar challenge but with added complexity. If a single gene is likened to an individual, then the collection of genes sharing common ancestry, typically performing the same role in different organisms, would be analogous to a species. Genes that are related in this way are commonly referred to as "orthologs" (*1*). Higher levels of gene or protein classification, such as families, subfamilies, and superfamilies, create a hierarchy in molecular taxonomy (*2*). Just what constitutes gene classification criteria can be uncertain in practice. This situation is made much more uncertain by the existence of nonorthologous relationships. Multiple proteins resulting from gene duplications within an organism are termed "paralogs." Paralogous relationships have been known for several decades: α-globin, β-globin, and myoglobin are classical examples of paralogs that arose from duplications of ancestral globin genes in the vertebrate lineage (*3*). In recent years, with the explosive increase in available sequence data, we have become aware of the richness of paralogous relationships in all organisms. We now realize that protein building blocks, or "modules," have duplicated and evolved in complex ways through a variety of gene-rearrangement mechanisms (*4*). As a result, composite proteins consisting of multiple modules ("chimeras") constitute a large proportion of the protein complement of an organism. The complexity that results from so many paralogous and chimeric relationships presents a daunting challenge for classification. Meeting the challenge unites sequence with biological information.

Like taxa, which reflect common ancestry but can also be used to infer common function, gene families have been of tremendous importance for understanding gene and protein function. Nearly all biological disciplines have profited from discoveries of family relationships. Such discoveries have reemphasized the importance of model systems in biology. For example, the sequencing of *Drosophila Ultrabithorax* and *Antennapedia* selector genes controlling segment identity delineated a shared homeobox module; this led to the discovery and intense study of related *HOX* genes in vertebrates and other organisms that are thought to play key roles in determining developmental fates (*5*). This example illustrates an increasingly popular paradigm in molecular genetics: Rather than proceeding from a phenotype to the isolation of a new gene, an investigator begins with the sequence of a key gene and searches for homologous genes in an organism of interest, preferably by scrutinizing the sequence databanks (*6*). Experimental data accumulated for the homologous (orthologous or paralogous) gene, when integrated with insights from gene family relationships, can accelerate our understanding of biological processes and our ability to rationally engineer genes.

Not just functional, but also structural inferences made from protein sequence alignments have been valuable to biologists. When a structure is known for one sequence, and another can be aligned with it, the unknown backbone structure can be predicted with confidence. In the case of homeoboxes, the high level of inferred structural similarity has guided site-directed modification of this DNA-binding domain for homeoboxes other than the structural archetype, and this situation holds for ~30% of known protein sequences (*7*).

Motifs, Modules, and Chimeras

The smallest sequence units of protein families are termed "motifs," which are identified as highly similar regions in alignments of protein segments (*8*). Motifs can be as simple as the hexamer repeat unit that forms a left-handed parallel β-helix found in uridine 5′-diphosphate (UDP)–*N*-acetylglucosamine acyltransferase (*9*). Motifs are widely used to identify functional regions of proteins and, where they share common ancestry, are useful for family classification. The C_2H_2 zinc finger DNA-binding motif, which is illustrated in the accompanying chart, defines the largest known family. By virtue of forming a contiguous independently folded structure, the finger is itself a module, whose small size of 21 to 26 amino acids is attributable to a zinc cation, which holds together two cysteine and two histidine residues from either end of the module. The larger homeobox module consists of a ~60–amino acid motif also involved in binding DNA. More typically, modules consist of multiple motifs, which form the structural core of proteins. Motifs contributing to a structural core can be widely separated within the primary sequence, as illustrated by the "HIGH" and "KMSKS" motifs of the Class I aminoacyl tRNA synthetases, which are hundreds of amino acids apart (*10*). Enzyme active site residues, which are usually highly conserved, are often found within motifs.

Motifs may reflect either common ancestry or convergence from independent origins. In either case, identification of motifs can be important for drawing structural and functional inferences. For example, the common "P-loop" motif is present in nucleotide-binding domains from families as diverse as kinesin motor proteins and adenosine 5′-triphosphate (ATP)–binding cassette (ABC) transporters, which are depicted in the accompanying chart. Despite the

S. Henikoff is at the Fred Hutchinson Cancer Research Center and Howard Hughes Medical Institute, Seattle, WA 98109–1024, USA. E. A. Greene and S. Pietrokovski are at the Fred Hutchinson Cancer Research Center, Seattle, WA 98109–1024, USA. P. Bork is at the European Molecular Biology Laboratory, 69012 Heidelberg, Germany, and Max-Delbrueck-Center for Molecular Medicine, 13122 Berlin-Buch, Germany. T. K. Attwood is in the Department of Biochemistry and Molecular Biology, University College London, London WC1E 6BT, UK. L. Hood is in the Department of Molecular Biotechnology, University of Washington, Seattle, WA 98195, USA.

*To whom correspondence should be addressed.

lack of a known structure for any ATP-binding cassette, the presence of a P-loop predicts the site of ATP binding in the transporter complex.

Modules are composed of single or multiple motifs. As the fundamental units of protein structure and function, modules are most useful for protein classification. Modules frequently display different connectivity relationships (Fig. 1, A to F), as illustrated by the kinesins and ABC transporters. The kinesin motor domain can be at either end of a polypeptide chain that includes a coiled-coil region and a cargo domain (*11*). ABC transporters are four-domain proteins consisting of two unrelated modules, a pair of ATP-binding cassettes, and a pair of integral membrane modules, which can be connected in different ways (*12*) (Fig. 1C).

Dispersal of Protein Building Blocks

Family relationships evolve over long periods of time by speciation and by sequence duplications fixed in genomes. Even the most recently evolved family relationships are still so ancient that the events that gave rise to paralogs and chimeras in modern genomes cannot be directly observed. However, enough is known about genomic-rearrangement mechanisms that some inferences can be drawn. Chromosomes evolve by transposition of mobile elements; by gross rearrangements such as inversions, translocations, deletions, and duplications; by homologous recombination; and by slippage of DNA polymerases during replication. It is likely that all of these mechanisms have contributed to the proliferation and dispersal of protein building blocks. Modules present in larger proteins, including homeobox modules, might have dispersed by transposition. Tandemly repeated modules, including the C_2H_2 zinc fingers and many examples of extracellular modules, most likely arose by recombinational mechanisms, such as unequal crossing-over and gene conversion (Fig. 1, A and E).

Multiple eukaryotic biosynthetic enzymes, especially those in the purine and pyrimidine pathways, are sometimes found together within a single polypeptide, unlike their separately encoded bacterial orthologs (*13*). For example, vertebrates have a multienzyme polypeptide for GAR synthetase, AIR synthetase, and GAR transformylase (GARS-AIRS-GART) (*14*). In insects, the polypeptide appears as GARS-$(AIRS)_2$-GART; in yeast, GARS-AIRS is encoded

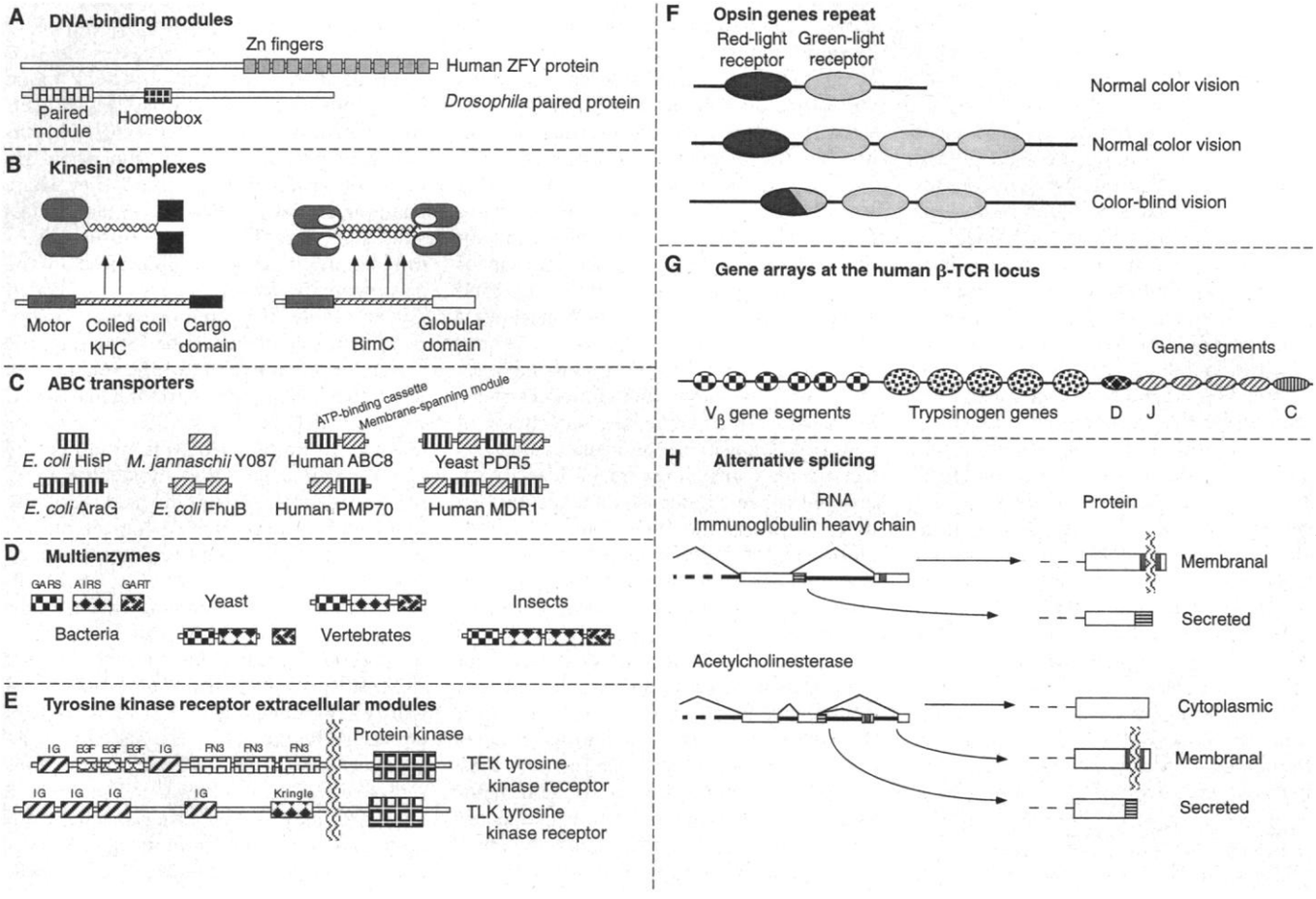

Fig. 1. Schematic representations of various building block arrangements described in the text. (**A**) Simple building blocks in DNA-binding proteins. The human ZFY protein contains 13 tandemly repeated zinc finger modules, and the *Drosophila* paired protein contains a paired box and a homeobox. (**B**) Subfamily relationships as predictors of quaternary structure: dimeric kinesin heavy chain (KHC) and tetrameric BimC protein complexes. (**C**) ABC transporters display different connectivities of two subunit pairs. Other examples of circular permutation have been recently reviewed (*54*). (**D**) Organism-specific fusion and duplication of purine biosynthetic pathway orthologs to GARS, AIRS, and GART. (**E**) Diverse modules are found in the extracellular portion of protein tyrosine kinases. (**F**) Humans are polymorphic for duplications and deletions within the opsin tandem cluster of long-wavelength genes. (**G**) T cell receptor (TCR) genes are interrupted by clusters of β-trypsinogen genes. (**H**) Alternative processing produces membrane-bound, secreted or intracellular forms of antibodies (or both), and acetylcholinesterases.

separately from GART; and in bacteria, GARS, AIRS, and GART are all encoded separately (Fig. 1D). The sites of fusion may correspond to introns, suggesting that chromosomal rearrangements have fused transcription units within introns. In other cases, fusions might have occurred in exons, or intron loss might have erased evidence of intron-mediated fusion (*15*). Regardless of mechanism, the fusion of transcription units is likely to have contributed to combining of protein building blocks in both eukaryotes and prokaryotes.

The mechanisms that gave rise to the dispersal of paralogous proteins within genomes are also diverse and frequently uncertain. The rhodopsin-like guanosine 5′-triphosphate (GTP)–binding protein (G protein)–coupled receptors illustrate multiple dispersal patterns (*16*). This family includes hormone, neurotransmitter, light, and olfactory receptors that are distinguished from one another by both sequence and functional differences. Remarkably, there are several hundred human olfactory receptor (OR) genes present in a dozen or so tandem clusters on several chromosomes (*17*). A cluster of three OR genes and an OR pseudogene fused to a different OR gene is thought to have arisen from disparate events, including recombinations between repeats flanking OR genes and a fusion by nonhomologous deletion (*18*).

Tandem gene clusters are sometimes interrupted by paralogous members of other gene families. For example, intercalated between repeated coding elements of the human β T cell receptor (TCR) locus are five trypsinogen genes in inverted orientation (*19*) (Fig. 1F). This complex arrangement of genes is likely to be of functional significance, as it is also found in mice and chickens.

Many paralogous relationships might be the consequence of whole-genome duplications. Ancient tetraploidization events in eukaryotes have been obscured by subsequent divergence, interchromosomal duplications, and other rearrangements but can be detected by careful analysis of genomic sequence. For example, it has been proposed that the *Saccharomyces* genome underwent a whole-genome duplication, and that 13% of *Saccharomyces cerevisiae* genes trace their lineage to this event (*20*). Tetraploidization events are common among higher plants; for example, the wheat genome consists of three copies of an ancestral grass genome. The human genome is thought to be the product of multiple tetraploidization events that occurred during chordate evolution (*5*). As a result, we have four copies of many genes or gene families, including four *HOX* gene clusters comparable to a single set of *HOX* genes in invertebrates. Enough time has passed since these putative tetraploidization events that vertebrate *HOX* genes have acquired distinguishable functions.

Selection for Diversity

The acquisition of a new specificity or a modified function after a gene-duplication event is often detectable by protein sequence comparison. For example, α-globins are more closely related to one another than they are to any β-globin. Maintenance of an acquired function over long evolutionary intervals can contribute greatly to the understanding of gene specificity. For example, sequence differences are sufficient to distinguish among tRNA synthetases that charge different amino acids, even though they belong to the same ancestral family (*21*). The kinesin motor domains provide another example, where relationships within a family are predictors for quaternary structural features: BimC motor domains are found in bipolar complexes, rather than in asymmetric complexes characteristic of other kinesin motors (*22*) (Fig. 1B). Comparisons should be interpreted with caution, especially when sequences from very distant organisms are compared; apparent subfamily relationships will not always reflect shared function. Furthermore, similar functions can arise in separate subfamilies. For example, among the ABC transporters, iron uptake is a function of members of two distinct subfamilies (*23*).

Relatively recent duplication events are sometimes responsible for diversity in molecular recognition. Tandem duplication of immunoglobulin (Ig) and TCR variable, joining, and diversity gene segments is the prototypical example, and special mechanisms of somatic DNA rearrangement and mutation further diversify antibody and TCR specificity. Among the rhodopsin-like G protein–coupled receptors, different olfactory receptors are thought to recognize different odorants, and different opsins are stimulated by different wavelengths of light. Long- and short-wavelength opsin genes diverged from one another early in vertebrate evolution (*24*). The opsins of the human visual system are present in a cluster on the X chromosome, with the long-wavelength opsins, sensitive to red and green light, constituting a tandem repeat with 98% sequence identity (Fig. 1F). Remarkably, the number of long-wavelength genes is polymorphic, a consequence of unequal crossing-over events that have occurred during human evolution. People with "normal" vision have a single red gene and one to three green genes. People who are red-green colorblind have lost a long-wavelength gene through a fusion of red and green tandem copies.

The products of gene duplication can act combinatorially and so further increase diversity. A response to a single antigen generally stimulates the proliferation of different B cells, each expressing a single antibody; the combination of different light and heavy chains provides heightened specificity to antigen. For olfaction, the stimulation of multiple olfactory receptors by their different odorants allows complex mixtures to be recognized. Our ability to recognize a full spectrum of colors with only three types of opsins is another example of the integration of multiple sensory inputs that have originated from duplicated building blocks.

Duplication of building blocks within a protein also results in generation of diversity during evolution. Each C_2H_2 zinc finger in a DNA-binding protein can recognize a 3–base pair motif, and in combination, multiple zinc fingers can mediate the binding to more complex DNA recognition sites (*25*). Combinatorial recognition by tandem zinc fingers has been exploited by researchers for designing new DNA-binding proteins (*26*). Combinations of unrelated modules have also broadened the spectrum of DNA-binding recognition, such as the presence of a paired box and a homeobox module in proteins related to *Drosophila paired* (*27*) (Fig. 1A). Extracellular proteins are notable for containing combinations of multicopy tandem arrays of different modules. The extracellular portion of the receptor tyrosine-specific class of protein kinases contains an astonishing variety of modules representing different families. For example, *trk*-like kinases have one kringle and four Ig modules, whereas *tek*-related proteins have three fibronectin III, three epidermal growth factor (EGF), and two Ig modules in their extracellular NH_2-terminal portions (*28*) (Fig. 1E). These extracellular modules can acquire diverse functions in different proteins. For example, some EGF modules bind to specific receptors, whereas others mediate interactions through calcium binding; the latter sometimes form long, rodlike structures composed of tandem module arrays (*29*).

Unlike germ-line processes that recombine gene segments during evolution, alternative messenger RNA (mRNA) processing can increase the diversity of proteins in the soma. For example, an alternative polyadenylation site within an intron of the Ig heavy-chain gene allows a switch from the synthesis of a membrane-bound receptor to a secreted antibody (*30*) (Fig. 1H). Acetylcholinesterase provides an example of alternative 3′ splice site selection accomplishing a comparable task; the choice of one terminal

exon leads to the synthesis of a glycophospholipid membrane anchor, the choice of the other to a cytoplasmic form, and lack of splicing to a secreted form of the enzyme (*31*).

Why Are Some Families So Large?

The accompanying chart provides information on the distribution of selected building blocks in model organisms. For organisms with completely determined genomic sequences, we can ask why some families are more successful than others. In *Escherichia coli*, the ABC transporters are the most common proteins encoded; this might reflect a flexible diet, which requires the uptake of diverse nutrients (*12*). It is likely that the much smaller number of ABC transporters in *Mycoplasma genitalium* and *Methanococcus jannaschii* reflect more limited diets. In general, paralogs account for half of all *E. coli* genes (*32*), which is high compared to the fractions found for smaller bacterial genomes, such as *Haemophilus influenzae*, where one-third of all genes are paralogs (*33*, *34*). Much of this difference is attributable to the more diverse nutritional and metabolic requirements of *E. coli* (*34*).

For organisms that have not yet been fully sequenced, it is necessary to extrapolate from samples of available sequences. For example, on the basis of finding only eight homeobox genes in *S. cerevisiae*, extrapolation predicts about 20 each in flies and worms, which are estimated to have two to three times as many genes (see accompanying chart). The fact that there are already about 60 genes reported in each of these two complex multicellular organisms demonstrates that homeobox genes have more successfully proliferated in animals than in a yeast. Although the number from *Drosophila melanogaster* is based on only ~10% of its genome, we predict that most of its homeobox genes have already been identified, and the final number will not be much greater than the number in *Caenorhabditis elegans* (which has nearly the same sized genome, ~70% of which is already sequenced). Such disproportionate representation of particular families is both a manifestation of their intense interest to researchers and of the ability to obtain these members by hybridization and amplification methods. Not all modules are as amenable to this approach as are the homeoboxes, which are especially highly conserved; to an increasing extent, partial complementary DNA (cDNA) sequencing projects are being used to identify coding sequences for gene families of interest (*35*). Many other gene families, such as the globins and the immunoglobulins, are disproportionately represented in collections of human sequences because they are important for human health (Table 1).

Even for the whole-genomic sequences that are currently available, the final size of known families is uncertain. Distant homologs may lie just beyond the horizon of current homology-detection methods. However, the introduction of improved methodology continues unabated, and this has led to the discovery of new family members and interfamily relationships. Moreover, the increasing size of a family can be exploited by multiple sequence-based methods to identify additional members (*36*). For example, 12 years ago, the similarity between opsin genes from human and fly was barely at the level of detection (*37*), yet today, the opsins are recognized as a closely related cluster within the rhodopsin-like G protein–coupled receptors (see accompanying chart). Most importantly, the accumulation of experimental evidence concerning gene or protein function or protein structure will provide insights that can be used to deduce possible family relationships that would not be compelling by sequence comparison methods alone.

Phylogenetic Distribution of Families

Size of a family within an organism is only one measure of success. Another is presence of a family in diverse organisms. Some families are successful at both, such as the ABC transporter family, which is not only one of the largest families overall (Table 1), but also appears to be present in all organisms. Most other families that are so widely distributed show much less proliferation within organisms. These include metabolic enzymes and components of the translational apparatus, which have only a few close paralogs (*38*). These families show a similar distribution to that of the GARS module in the table of the accompanying chart (*39*).

The chymotrypsin family of serine proteases is notable in being both ancient and large (Table 1), but the extreme proliferation appears to be confined to eukaryotes; only rarely are family members found in bacteria. This raises the possibility that other families that appear to be confined to certain branches of the tree of life are actually more ancient, but that they have simply become extinct in other lineages, or that a relationship has gone undetected. The latter is the case for eukaryotic tubulin and bacterial FtsZ, both of which use GTP for polymerization to form similar intracellular fibers and are believed to be ancestrally related (*40*). This relationship was not detected by pairwise sequence comparisons, but rather by recognition of a tubulin motif in FtsZ. Potentially homologous proteins have also been identified by structure determination, such as the detection of similar folds for kinesin and myosin motor proteins (*41*).

Given the extreme uncertainty in tracing the birth of a family, we nevertheless recognize that some families have proliferated to a remarkable extent in certain phyla. GAL4 transcriptional regulators, one of the largest families in yeast, have been found only in fungi (see accompanying chart). The EGF module, present in about 1% of human proteins, has been described only in animals (Table 1). The Ig module, which is found in more than 200 proteins in addition to all of the immune receptors (antibodies, TCRs, class I and II families of the major histocompatibility complex), is involved in diverse cell surface recognition phenomena in multicellular organisms (*42*). The Ig module has also successfully proliferated within proteins: A total of 244 copies of Ig and

Table 1. The largest protein families. The sources for these numbers of modules are Pfam (PF) or Prints (PR). GPCR, G protein–coupled receptor; LDL, low density lipoprotein.

Family	Source	Modules in SwissProt	Found where?
C_2H_2 zinc fingers	PF00096	1826	Eukaryotes, archaea
Immunoglobulin module	PF00047	1351	Animals
Protein (Ser/Thr/Tyr) kinases	PF00069	928	All kingdoms
EGF-like domain	PF00008	854	Animals
EF-hand (Ca binding)	PF00036	790	Animals
Globins	PF00042	699	Eukaryotes, bacteria
GPCR-rhodopsin	PF00001	597	Animals
Fibronectin type III	PF00041	514	Eukaryotes, bacteria
Chymotrypsins	PR00722	464	Eukaryotes, bacteria
Homeodomain	PF00046	453	Eukaryotes
ABC cassette	PF00005	373	All kingdoms
Sushi domain	PF00084	343	Animals
RNA-binding domain	PF00076	331	Eukaryotes
Ankrin repeat	PF00023	330	Eukaryotes
RuBisCo large subunit	PF00016	319	Plants, bacteria
LDL receptor A	PF00057	309	Animals

distantly related fibronectin III modules account for most of the 30,000-residue muscle titin protein (*43*). The success of the ~100–amino acid Ig module is attributable to its potential to undergo diversification in the presence of a highly conserved structural framework, its protease resistance in the folded form, and its ability to readily form homo- and heterodimers through multiple interacting surfaces, so that it is especially suitable for mediating cell-cell interactions.

Proliferation of one family might have occurred at the expense of others. The distribution of protein kinases is suggestive, in that the family consisting of serine-, threonine-, and tyrosine-specific enzymes is hugely successful only in eukaryotes, but is poorly represented in bacteria (see accompanying chart). Conversely, the family of histidine-specific protein kinases is highly successful in *E. coli* and other bacteria, but is relatively rare in eukaryotes. In such situations, we must also consider the possibility that these families are recent arrivals in some organisms, having been transferred horizontally between kingdoms. Horizontal transfers are difficult to document unless there are conspicuous anomalies evident from molecular phylogenetic analyses. Such anomalies have indicated numerous horizontal transfers of mariner transposases between diverse animals (*44*), as well as transfer of the fibronectin III module from a eukaryote to a bacterium (*45*).

The establishment, proliferation, or extinction of a protein family in a lineage may coincide with a functional innovation during evolution. For example, actins, tubulins, and motors such as kinesins are found only where there is a cytoskeleton, as though the evolution of these proteins was coordinate with the appearance of the cytoskeleton in eukaryotes. In bacteria, σ factors regulate transcriptional initiation, in contrast to eukaryotes and archaea, which use a different system (*46*). This difference suggests that either the σ factor system coincided with the appearance of bacteria or that it was lost in the eukaryotic-archaea lineage.

Interspersed Genomewide Repeats

Analysis of whole-genomic sequences definitively demonstrates that coding regions of genes dominate the prokaryotic genome (*38*). In contrast, complex eukaryotic genomes are dominated by noncoding sequences. Families of repeats derived from transposable elements constitute a major portion of these eukaryotic genomes, far exceeding exons in the proportion of the genome devoted to them (*47*, *48*). Transposition can occur by reverse transcription of an RNA intermediate or by excision and reintegration of DNA itself (DNA transposition). These elements fall into four categories: short interspersed nuclear elements (SINEs), long dispersed nuclear elements (LINEs), long-terminal repeat (LTR) retrovirus-like elements, and DNA transposons (Fig. 2). In the human, there are ~1,100,000 Alu sequences (a SINE) and 590,000 Line1 sequences (a LINE). It is impressive that Line1 occupies an order of magnitude more of our genome than all of our gene-coding sequences combined. Furthermore, with improved techniques for identifying degraded repeat sequences, perhaps 50% of our genome and an even higher fraction of the mouse genome will be found to consist of genomewide repeats. Much of the nonassigned genome sequences might be composed of interspersed repeats degraded to the point that they are no longer recognizable.

Vertebrate chromosomes have large-scale mosaic structures, or isochores, often with distinct ratios of G+C nucleotides, repeat content, and gene density (*49*). The human contigs in Table 2 represent high (class II major histocompatibility locus)–, medium (TCR)–, and low (metabolic glutamate receptor 8)–gene density regions. Low–gene density loci are A+T- and Line1-rich, whereas high–gene density loci are G+C- and Alu-rich (*47*, *49*). The A+T-rich isochores, in general, contain longer genes.

The repeats may have at least three important functional and evolutionary roles. First, some may evolve to become the regulatory regions of genes expressed in a tissue-specific manner (*50*). Second, repeats play an important role in refashioning the genomic architecture by facilitating homologous recombination, translocations, and perhaps gene conversions. And third, repeats have been implicated in epigenetic phenomena, such as parental imprinting and position-effect variegation (*51*). Because the ages of repeats can be determined by species comparisons, they can serve as valuable time markers for unraveling the complexities of molecular archaeology in complex gene loci such as the TCR genes.

Table 2. Content of long contguous stretches of DNA sequence in selected human and mouse gene regions. Data are from the Leroy Hood laboratory.

Region	Contig length (bp)	GC (%)	mRNA (%)	Interspersed repeats (%)	Line1 (%)	Alu or B1/B2 (SINES) (%)
Human TCRα	1,071,650	40	4.0	35	16	8
Mouse TCRα	228,654	41	1.5	33	22	2.4
Human TCRβ	684,973	42	4.6	30	14	5
Human TCR on chromosome 9	216,293	41	1.7	45	23	9
Mouse TCRβ	700,960	40	3.8	43	32	2
Human MHC class III	299,287	52	16.8	30.5	6.7	17

Prospects

There is good news and bad news for gene taxonomists. The good news is that the number of identified protein families has been increasing only slowly with the rapid increase in new sequence data and is expected to level off. The bad news is that family relationships are so complex that we cannot use any simple hierarchical scheme to make the data easily understandable. Nevertheless, as more is learned from model organisms about individual modules, their presence in any protein of interest adds potential insight into its function and guides experiments, which is good news for biologists. Gene taxonomists have learned by now to cope with complexity in family relationships, and currently several classification systems are used to construct the different databases

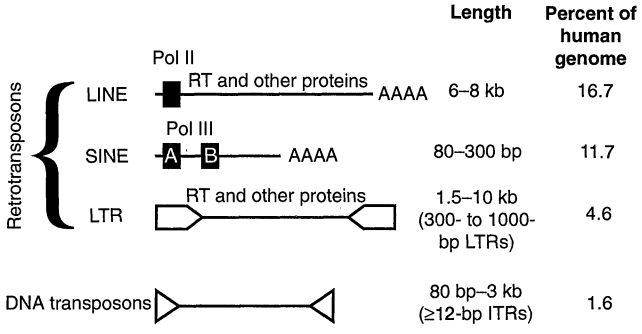

Fig. 2. Schematic representation of the types of transposable elements that have produced high-copy number human interspersed repeats. The shaded boxes denote internal promoter sites; names inside the bracket indicate that only autonomous elements code for these proteins. LTR, long-terminal repeat; ITR, inverted-terminal repeat; RT, reverse transcriptase. [Adapted from (*47*) on the basis of 7051 kb of human sequence]

listed in the accompanying chart. In fact, the task of classification is made easier for gene taxonomists than for Linnaean taxonomists because sequence similarity is a precisely defined metric for establishing relatedness. This metric makes possible automated and computer-assisted classifications of genes. Much more difficult is the task of enriching the databases of genes and families with insights obtained from experiments.

To some extent, computer-based tools can be applied to the task of connecting genes and families with information about them. Organism-specific databases and retrieval tools such as the National Center for Biotechnology Information's Entrez allow biologists to rapidly obtain needed information from the World Wide Web. However, insight cannot be automated, and computer-based tools that go beyond sophisticated retrieval methods may not be the solution. One problem is that generalized databases are too constraining to allow more than minimal documentation of individual protein families. Another problem is that the literature pertaining to a single family can be so vast that only an expert devoted to that family can master it. Fortunately, a number of biologists interested in particular families have begun to exploit the Web to provide the kind of rich information that can be used to gain insight into function. At a single family Web site, participation can be distributed among multiple laboratories, and information can be continually updated and integrated (*52*). Furthermore, new Web sites are developed on the basis of existing sites. There are currently five Web sites dedicated to different nuclear hormone receptors spawned from the Nuclear Receptor Resource, and the Myosin Web site was spawned from the Kinesin Web site (*53*). An organized effort to develop such sites is in progress (see http://proweb.org for information on participating).

We have focused here and in the accompanying chart primarily on large and well-studied families. But to truly understand a biological system, we will need to understand the interaction of all individual components. Some of these components will not be immediately classifiable. Eventually, detectable homologs for most of these "orphans" will be discovered in genome-sequencing projects. As a result, new family relationships will become delineated that are useful for identifying critical regions and guiding experimental work. This situation is most evident in an organism such as *M. jannaschii*, for which a large fraction of proteins are as yet unclassified orphans, but to a lesser extent it is true for all major phyla. The identification and classification of new protein families and the deep insights that result should continue well into the next millennium.

REFERENCES AND NOTES

1. W. Fitch, *Syst. Zool.* **19**, 99 (1970). Orthologs can only be determined definitively with a complete inventory of the genes in an organism. See R. L. Tatusov, E. V. Koonin, D. J. Lipman, *Science* **278**, 631 (1997).
2. We use the term "family" generically to describe any collection of genes or proteins that are presumed to share common ancestry.
3. V. M. Ingram, *Hemoglobins in Genetics and Evolution* (Columbia Univ. Press, New York, 1963).
4. Modules are contiguous in sequence, whereas structural domains are independently folded units that need not be contiguous [L. Patthy, *Cell* **41**, 657 (1985); S. Henikoff, J. C. Wallace, J. P. Brown, *Methods Enzymol.* **183**, 111 (1990); P. Green *et al.*, *Science* **259**, 1711 (1993); R. F. Doolittle, *Annu. Rev. Biochem.* **64**, 287 (1995)].
5. W. J. Gehring and Y. Hiromi, *Annu. Rev. Genet.* **20**, 147 (1986). F. H. Ruddle *et al.*, *ibid.* **28**, 423 (1994).
6. R. F. Doolittle, *Of URFs and ORFs: A Primer on How to Analyze Derived Amino Acid Sequences* (University Science Books, Mill Valley, CA, 1987).
7. C. Orengo, *Curr. Opin. Struct. Biol.* **4**, 429 (1994); R. Schneider, A. de Daruvar, C. Sander, *Nucleic Acids Res.* **25**, 226 (1997).
8. The term "motif" has different interpretations. See P. Bork and E. V. Koonin, *Curr. Opin. Struct. Biol.* **6**, 366 (1996).
9. C. R. H. Raetz and S. L. Roderick, *Science* **270**, 997 (1995).
10. P. Schimmel, *Trends Biochem. Sci.* **16**, 1 (1991).
11. J. D. Moore and S. A. Endow, *Bioessays* **18**, 207 (1996).
12. M. Dean and R. Allikmets, *Curr. Opin. Genet. Dev.* **5**, 779 (1995).
13. J. N. Davidson *et al.*, *Bioessays* **15**, 157 (1993). J. N. Davidson and M. L. Peterson, *Trends Genet.* **13**, 281 (1997).
14. GARS (glycinamide ribonucleotide synthetase), AIRS (aminoimidazole ribonucleotide synthetase), and GART (glycinamide ribonucleotide transformylase).
15. A. Rzhetsky *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 6820 (1997); S. J. de Souza *et al.*, *ibid.* **93**, 14632 (1996).
16. G protein–coupled receptors are a "clan," which includes proteins that may not be ancestrally related to rhodopsin.
17. N. Ben-Arie *et al.*, *Hum. Mol. Genet.* **3**, 229 (1993). R. R. Reed, *Cold Spring Harbor Symp. Quant. Biol.* **57**, 501 (1992).
18. G. Glusman *et al.*, *Genomics* **37**, 147 (1996).
19. L. Rowen, B. F. Koop, L. Hood, *Science* **272**, 1755 (1996).
20. K. H. Wolffe and D. C. Shields, *Nature* **387**, 708 (1997).
21. R. Wetzel, *J. Mol. Evol.* **40**, 545 (1995).
22. A. S. Kashina *et al.*, *Nature* **379**, 270 (1996).
23. A. Angerer, S. Gaisser, V. Braun, *J. Bacteriol.* **172**, 572 (1990).
24. J. Nathans *et al.*, *Annu. Rev. Genet.* **26**, 403 (1992).
25. Y. Choo and A. Klug, *Curr. Opin. Struct. Biol.* **7**, 117 (1997).
26. ———, *Curr. Opin. Biotechnol.* **6**, 431 (1995).
27. D. Bopp *et al.*, *Cell* **47**, 1033 (1986).
28. P. Maslakowski and R. D. Carroll, *J. Biol. Chem.* **267**, 26181 (1992); J. Partanen *et al.*, *Mol. Cell. Biol.* **12**, 1698 (1992).
29. P. Bork *et al.*, *Q. Rev. Biophys.* **29**, 119 (1996).
30. J. Rogers *et al.*, *Cell* **20**, 303 (1980).
31. Y. Li, S. Camp, P. Taylor, *J. Biol. Chem.* **268**, 5790 (1993).
32. B. Labedan and M. Riley, *Mol. Biol. Evol.* **12**, 980 (1995).
33. S. E. Brenner *et al.*, *Nature* **378**, 140 (1995).
34. R. L. Tatusov *et al.*, *Curr. Biol.* **6**, 279 (1996).
35. M. D. Adams *et al.*, *Science* **252**, 1651 (1991).
36. R. F. Doolittle, Ed., *Methods Enzymol.* **266** (1996).
37. J. E. O'Tousa *et al.*, *Cell* **40**, 839 (1985).
38. R. L. Tatusov *et al.*, in (*1*).
39. For the table in the accompanying chart, organism-specific counts were obtained for C_2H_2 zinc fingers (Pfam PF00096), homeodomains (Blocks BL00027), LysR transcription regulators (BL00044), TATA-binding protein repeat (BL00351), 7TM rhodopsdin-like receptors (Prints GPCRRHODOPSN), kinesin motors (BL00411), ATP-binding cassette (BL00211), DEAD/H helicases (PF00270), AAA modules (BL00674), hsp60s (BL00296), and hsp20s (BL01031) by MAST searches [T. L. Bailey and M. Gribskov, *J. Comp. Biol.* **4**, 45 (1997)] of OWL version 29.3 by use of position-specific scoring matrices from local multiple alignments [J. G. Henikoff and S. Henikoff, *Comput. Appl. Biosci.* **12**, 135 (1996)]. For GAL4 transcription regulators, Ser-, Thr-, Tyr-specific kinases, His-specific kinases, kringle extracellular domain, WW intracellular domain, BRCA1 COOH-terminal domain, and Calponin homology domain, profiles were constructed from multiple alignments and used to search an exhaustive protein database at the European Molecular Biology Laboratory, Heidelberg, Germany, with exclusion of redundant entries [P. Bork and T. J. Gibson, *Methods Enzymol.* **266**, 162 (1996)].
40. H. P. Erickson, *Cell* **80**, 367 (1995).
41. F. J. Kull *et al.*, *Nature* **380**, 550 (1996).
42. T. Hunkapiller and L. Hood, *Adv. Immunol.* **44**, 1 (1989).
43. S. Labeit and B. Kolmerer, *Science* **270**, 293 (1995).
44. H. M. Robertson, *Nature* **362**, 241 (1993); *J. Hered.* **88**, 195 (1997); ——— *et al.*, *Nature Gen.* **12**, 360 (1996).
45. P. Bork and R. F. Doolittle, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 8990 (1992).
46. D. Langer *et al.*, *ibid.* **92**, 5768 (1995).
47. A. F. A. Smit, *Curr. Opin. Genet. Dev.* **6**, 743 (1996).
48. P. SanMiguel *et al.*, *Science* **274**, 765 (1996); J. A. Yoder, C. P. Walsh, T. H. Bestor, *Trends Genet.* **13**, 335 (1997).
49. G. Bernardi, *Annu. Rev. Genet.* **29**, 445 (1995).
50. J. Brosius, *Science* **51**, 753 (1991); S. E. White, L. F. Habera, S. R. Wessler, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 11792 (1994); R. J. Britten, *ibid.* **93**, 9374 (1996).
51. S. Henikoff and M. A. Matzke, *Trends Genet.* **13**, 293 (1997); D. P. Barlow, *Science* **260**, 309 (1993).
52. S. Henikoff, S. A. Endow, E. A. Greene, *Trends Biochem. Sci.* **21**, 444 (1996).
53. Nuclear Receptor Resource, http://nrr.georgetown.edu/NRR/NRR.html; Kinesin Home Page, http://proweb.org/kinesin; Myosin Home Page, http://proweb.org/myosin.
54. Y. Lindqvist and G. Schneider, *Curr. Opin. Struct. Biol.* **7**, 422 (1997).

55. Supported by grants from NIH (GM29009) and U.S. Department of Energy (DE-FG03-97ER62382). S.P. is a Howard Hughes Medical Institute Fellow of the Life Sciences Research Foundation. T.K.A. is a Royal Society University Research Fellow. P.B. thanks J. Schultz and M. Huynen for helpful discussions.