SCIENCE

Building Gene Familie

Genom

Kinesins are eukaryotic motor proteins in which the module is a motor domain that uses A "walk" along microtubules. The motor domain is a single fold with multiple conserved region an ATP-binding motif. The two motor domains of dimeric kinesin proteins are each connecte coiled-coil stalk that holds the monomers together. At the other end of the stalk is a globular can be used for attachment to cargo. The motor domain can be at either end of the sequenc dle, and is highly conserved, whereas cargo domains are variable in sequence. Most kinesins are built from parallel dimers and are involved in vesicle transport or spindle movements. However, in

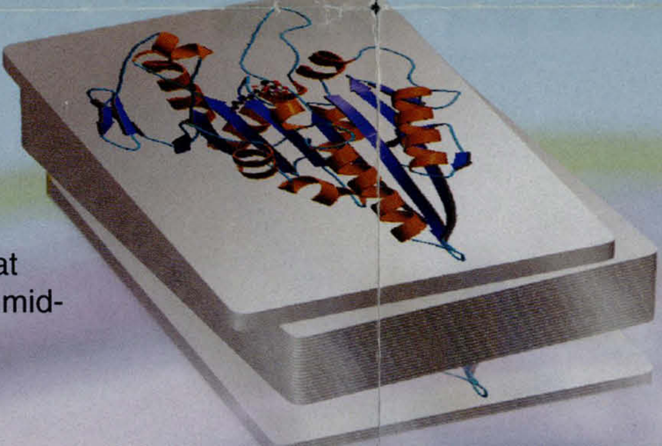
Genome sequencing projects and other large-scale efforts are generating hundreds of thousands of sequences of new proteins from diverse organisms. The task of discovering the structure and function of an unknown protein is aided by the fact that most new genes are related to other genes, and these relationships can often be detected via sequence similarity. Perhaps half of all known genes encode members of some 3000 major families. Family members share sequence and structural similarities, suggesting divergence from a common ancestor. Unlike proteins that are direct counterparts in different organisms, there can be many members of a gene family within one organism that carry out distinct, yet similar, functions. For the organism itself, the existence of gene families provides a way of generating diversity in function and specificity from a limited number of building blocks, which is essential for the evolutionary success of a genome. Within large eukaryotic genomes, gene family size varies tremendously, ranging from a unique member to thousands of members. Even smaller genomes harbor families that comprise several percent of their genome.

In the image, the tornado symbolizes the powerful forces that reorganize and disperse the building blocks. Near the bottom,

es

m e m a

at uses ATP to
ed regions including
connected to a
a globular domain that
sequence, or in the mid-



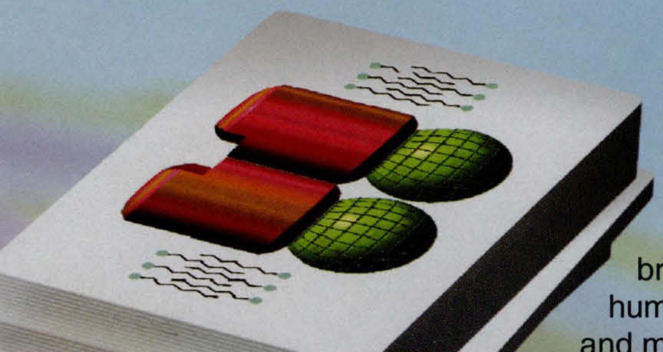
orts are
ew pro-
ring the
the fact
se rela-
Perhaps
0 major
similar-
. Unlike
s, there
rganism
rganism
generat-
mber of
success
mily size
to thou-
lies that

ces that
bottom,

card fragments represent building blocks of genes encoding emerging proteins, complete cards represent functional genes, and stacks of cards at the top represent gene families. Although there is a large element of randomness in this process, the tornado is under constraints. These are shown as icons (see below), which illustrate organizational features. These govern the evolution of protein modules from motifs, of complete proteins from modules, and of gene families from duplication of individual genes.

Motifs are short, conserved sequence regions. A module is a contiguous sequence segment that can consist of a single motif —▮— or multiple motifs in fixed order —▮▮▮—. A protein may result from a single module —▮— or concatenation of multiple, independent modules —▮▮▮—. The same module may also be repeated within a protein —▮▮▮— . Gene families can be tandemly duplicated —⬬⬬— or dispersed ⬬⬬ . Regions of chromosomes containing many genes can be duplicated and dispersed ⬬⬬ . These features are illustrated by the selected examples shown below, and further discussion can be found in the accompanying article in the 24 October, 1997 issue of SCIENCE.

aps 8



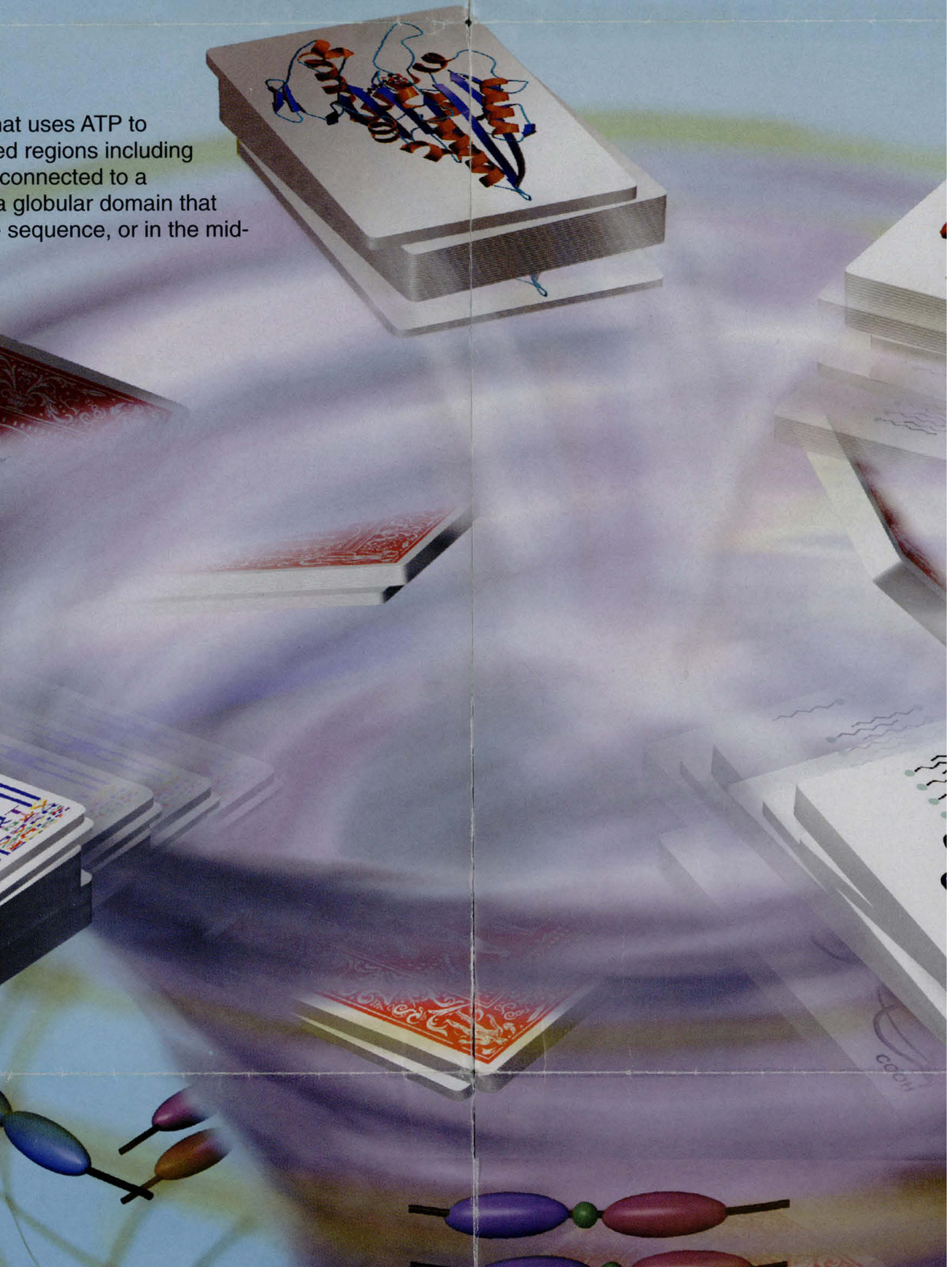
ABC proteins form a family characterized by a highly conserved ATP-binding cassette. The proteins are found in archaea, bacteria, and eukaryotes and almost all are transporters that import or export a diverse group of specific substrates across membranes. Family members are involved in human hereditary diseases, antigen processing, and multidrug resistance of protozoan parasites.

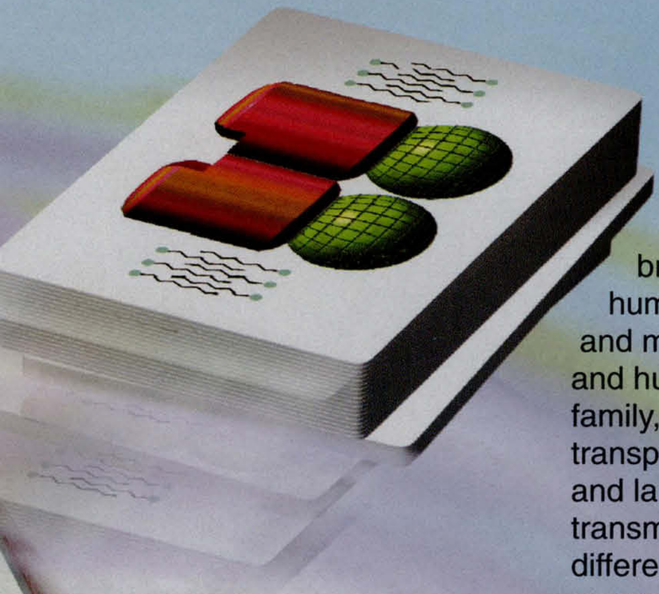
Kinesins are eukaryotic motor proteins in which the module is a motor domain that uses A "walk" along microtubules. The motor domain is a single fold with multiple conserved regions an ATP-binding motif. The two motor domains of dimeric kinesin proteins are each connecte coiled-coil stalk that holds the monomers together. At the other end of the stalk is a globular can be used for attachment to cargo. The motor domain can be at either end of the sequenc dle, and is highly conserved, whereas cargo domains are variable in sequence. Most kinesins are built from parallel dimers and are involved in vesicle transport or spindle movements. However, in one group antiparallel dimers form tetrameric motors that are proposed to facilitate microtubule sliding. In addition, some kinesins are thought to function as monomers, and others consist of heterodimeric kinesin proteins that, together with a third nonmotor subunit, function as heterotrimeric motors.

C_2H_2 Zinc finger proteins

The intimate relation between structure and sequence conservation is illustrated by the zinc finger motif, which is often involved in DNA binding and regulation of gene expression.The ribbon cartoon is a structural representation, and the logo below it is the corresponding sequence-based representation of a zinc finger. A sequence alignment based on zinc fingers from more than 100 proteins identifies the invariant and conserved positions. In each position of the logo, the degree of conservation is proportional to the height of each letter, and colors are assigned to reflect amino acid properties [http://www-lmmb.ncifcrf.gov/~toms/sequencelogo.html]. As seen in the structure, two cysteine and two histidine side chains coordinate the zinc. The fixed

hat uses ATP to
ed regions including
connected to a
a globular domain that
sequence, or in the mid-





ABC proteins

form a family characterized by a highly conserved ATP-binding cassette. The proteins are found in archaea, bacteria, and eukaryotes and almost all are transporters that import or export a diverse group of specific substrates across membranes. Family members are involved in human hereditary diseases, antigen processing, and multidrug resistance of protozoan parasites and human tumor cells. In different members of the family, ATP hydrolysis either regulates or energizes transport of molecules as diverse as small ions and large proteins. ABC transporters contain two transmembrane and two ATP-binding modules. In different proteins these modules can be encoded as separate polypeptides or can be fused together in different orders and combinations.

G protein–coupled receptors

(GPCRs) encompass a wide range of autocrine, paracrine, and endocrine processes. The rhodopsin-like GPCRs include families of hormone, neurotransmitter, odorant, and light receptors, all of which transduce extracellular signals through interaction with guanine nucleotide–binding (G) proteins. Although individual family members transduce different stimuli into disulation of gene expression.The ribbon cartoon is a structural representation, and the logo below it is the corresponding sequence-based representation of a zinc finger. A sequence alignment based on zinc fingers from more than 100 proteins identifies the invariant and conserved positions. In each position of the logo, the degree of conservation is proportional to the height of each letter, and colors are assigned to reflect amino acid properties [http://www-lmmb.ncifcrf.gov/~toms/sequencelogo.html]. As seen in the structure, two cysteine and two histidine side chains coordinate the zinc. The fixed number of amino acids between the internal cysteine and histidine reflects constraints on the backbone, whereas the side chains display widely varying degrees of conservation. For example, the first position after the internal cysteine has few constraints, and so a short stack of numerous different residues is seen. In contrast, the fourth position is preferentially aromatic (orange), resulting in a tall stack dominated by phenylalanine, tyrosine, and tryptophan. This conserved residue holds together the second β strand (cyan) and the α helix (magenta). Combining multiple zinc fingers within a protein allows binding to neighboring sites in DNA. The combinatorial flexibility that results may account for the extraordinary proliferation of this small module.

Databases of Protein Families

The information embedded in protein families is an essential resource for refined homology searching, identifying critical residues in a module, predicting secondary structure, deriving phylogeny of organisms and subfamily relationships, homology modeling of three-dimensional structure, and gene prediction. The BCM search launcher—http://kiwi.imgen.bcm.tmc.edu:8088/search-launcher/launcher.html—is an excellent starting point for exploring many of these methods. For more tools, a well-organized list of links can be found at http://www-biol.univ-mrs.fr/english/logligne.html. Many individual protein families are described in detail at dedicated sites; see http://www.proweb.org for a current guide, including information for researchers interested in establishing new sites.

Sequence Similarity

PROSITE—http://www.expasy.ch/sprot/prosite.html

Blocks—http://www.blocks.fhcrc.org/

Prints—http://www.biochem.ucl.ac.uk/bsm/dbbrowser/PRINTS/

ProDom—http://protein.toulouse.inra.fr/

Pfam—http://www.sanger.ac.uk/Pfam/
http://genome.wustl.edu/Pfam/

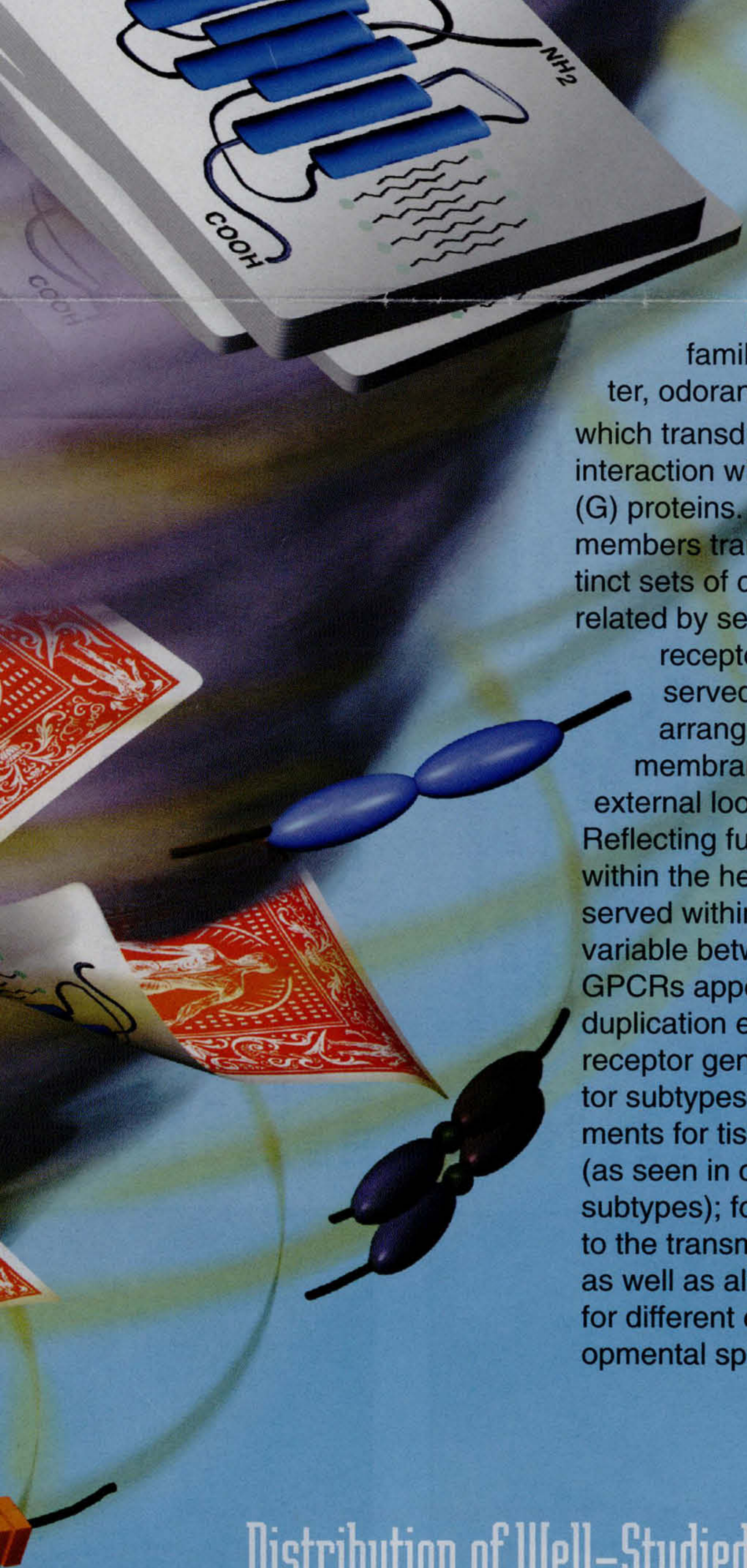
ProClass—http://diana.uthct.edu/proclass.html

Structure

CATH—http://www.biochem.ucl.ac.uk/bsm/cath/

SCOP—http://scop.mrc-lmb.cam.ac.uk/scop/

This table
methods im
cally meaning
modules are foun
function (e.g., regula
Alternatively, GARS
Protein counts were
the EMBL protein da



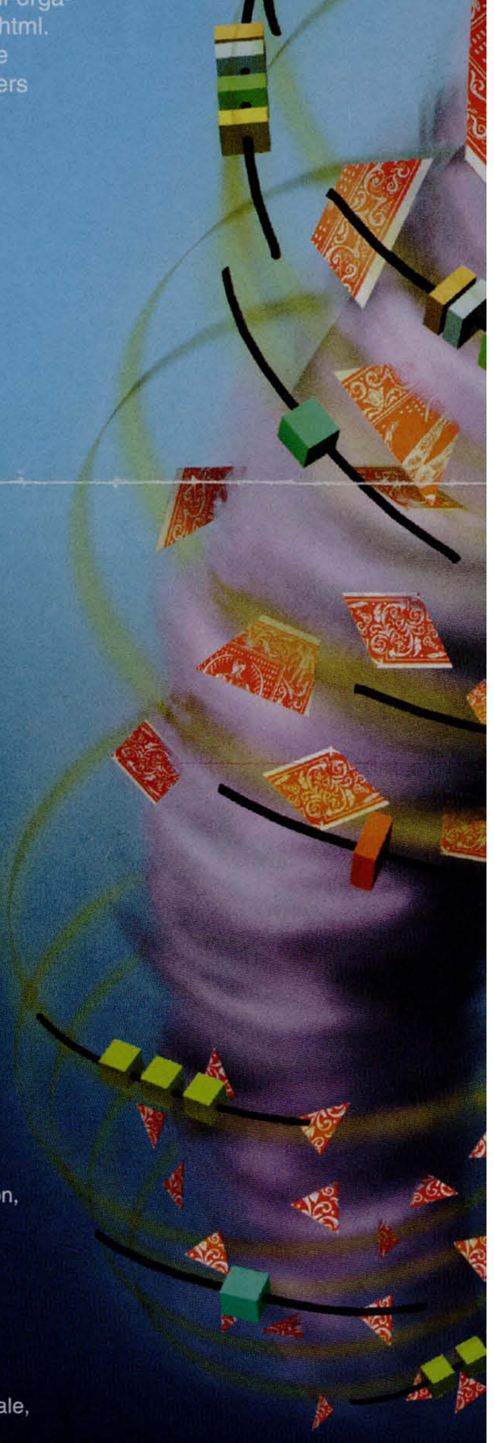
G protein–coupled receptors (GPCRs)

encompass a wide range of autocrine, paracrine, and endocrine processes. The rhodopsin-like GPCRs include families of hormone, neurotransmitter, odorant, and light receptors, all of which transduce extracellular signals through interaction with guanine nucleotide–binding (G) proteins. Although individual family members transduce different stimuli into distinct sets of cellular responses, they are related by sequences and shared folds. Each receptor is thought to fold into a conserved conformational switch, arranged as a bundle of seven transmembrane helices connecting three external loops to three cytoplasmic loops. Reflecting functional diversity, sequences within the helix bundle are somewhat conserved within receptor families, but are more variable between them. The rhodopsin-like GPCRs appear to have arisen via gene duplication events from a single ancient receptor gene. Evolutionary stability of receptor subtypes may be preserved by requirements for tissue and regional specialization (as seen in cardiac versus brain muscarinic subtypes); for conferring different sensitivities to the transmitter and to possible modulators, as well as alternative tranduction pathways for different effector systems; and for developmental specializations.

Distribution of Well–Studied Building Blocks

This table reflects the current state of building block classification; as new sequences become available and comparison
methods improve, the percentage of proteins classified into families increases and new families are delineated. A biologi-
ally meaningful comparison can be made by counting the occurrence of particular modules in different organisms. Some
ules are found in all organisms examined, whereas particular organisms may favor different modules for an analogous
(e.g., regulation of transcription). In the largest family, the C_2H_2 Zn finger proteins, the module has diverse functions.
vely, GARS is an example of a true ortholog, performing the identical enzymatic function in the different organisms shown.
counts were obtained from OWL v.29.3 and family counts from ProClass v.1.1. Modules were tallied by searching OWL or
3L protein database, with exclusion of redundant entries from the final count.

lent starting point for exploring many of these methods. For more tools, a well-organized list of links can be found at http://www-biol.univ-mrs.fr/english/logligne.html. Many individual protein families are described in detail at dedicated sites; see http://www.proweb.org for a current guide, including information for researchers interested in establishing new sites.



Sequence Similarity

PROSITE—http://www.expasy.ch/sprot/prosite.html

Blocks—http://www.blocks.fhcrc.org/

Prints—http://www.biochem.ucl.ac.uk/bsm/dbbrowser/PRINTS/

ProDom—http://protein.toulouse.inra.fr/

Pfam—http://www.sanger.ac.uk/Pfam/
http://genome.wustl.edu/Pfam/

ProClass—http://diana.uthct.edu/proclass.html

Structure

CATH—http://www.biochem.ucl.ac.uk/bsm/cath/

SCOP—http://scop.mrc-lmb.cam.ac.uk/scop/

LPFC—http://www-camis.stanford.edu/projects/helix/LPFC/

Metabolic Function

WIT—http://www.cme.msu.edu/wit/

KEGG—http://www.genome.ad.jp/kegg/

Credits:

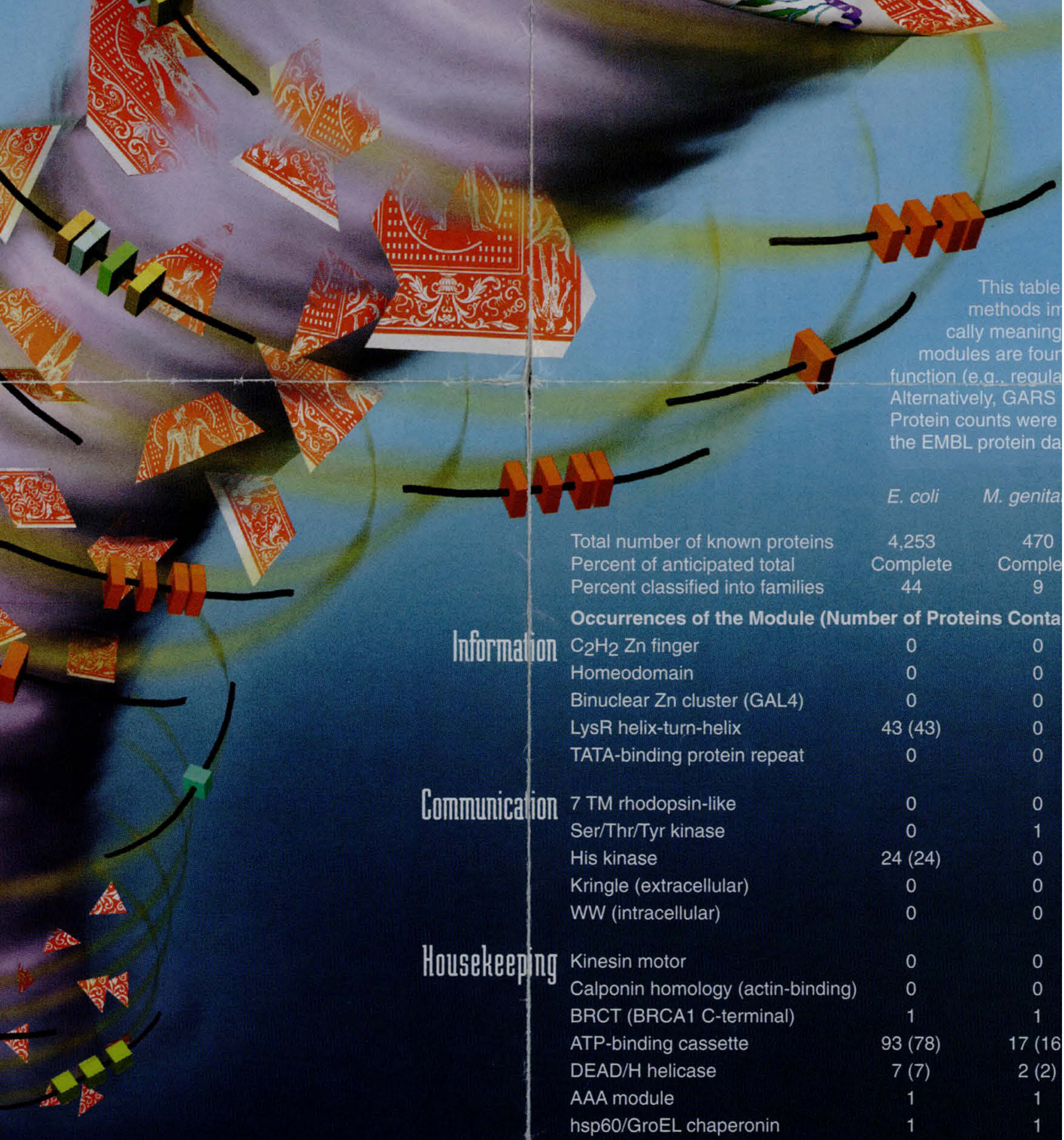
Science Coordinator: Barbara R. Jasny

Authors: Elizabeth A. Greene, Shmuel Pietrokovski, Steven Henikoff, Fred Hutchinson Cancer Research Center, Seattle, WA, USA; Peer Bork, EMBL (European Molecular Biology Laboratories, Heidelberg) and MDC (Max Delbrück Center, Berlin), Germany; Teresa K. Attwood, University College London, UK; Lee Hood, University of Washington, Seattle, WA,USA; Amos Bairoch, University of Geneva, Switzerland

Reviewers: Sean Eddy, Washington University School of Medicine, St. Louis, MO, USA; Eugene Koonin, National Center for Biotechnology Information, Bethesda, MD, USA; Jasper Rine, University of California, Berkeley, CA, USA. With special thanks to Q. Al-Awqati, H. Bourne, S. Endow, P. Millner, and R. Tjian

Design and Art Direction: Cynthia Faber Smith

Illustration: Steven Keller; kinesin motor image: F. J. Kull, E. Sablin, R. Fletterick, R. Vale, University of California, San Francisco, USA



This table
methods im
cally meaning
modules are foun
function (e.g., regula
Alternatively, GARS
Protein counts were
the EMBL protein da

		E. coli	*M. genita*
	Total number of known proteins	4,253	470
	Percent of anticipated total	Complete	Comple
	Percent classified into families	44	9
	Occurrences of the Module (Number of Proteins Conta		
Information	C_2H_2 Zn finger	0	0
	Homeodomain	0	0
	Binuclear Zn cluster (GAL4)	0	0
	LysR helix-turn-helix	43 (43)	0
	TATA-binding protein repeat	0	0
Communication	7 TM rhodopsin-like	0	0
	Ser/Thr/Tyr kinase	0	1
	His kinase	24 (24)	0
	Kringle (extracellular)	0	0
	WW (intracellular)	0	0
Housekeeping	Kinesin motor	0	0
	Calponin homology (actin-binding)	0	0
	BRCT (BRCA1 C-terminal)	1	1
	ATP-binding cassette	93 (78)	17 (16
	DEAD/H helicase	7 (7)	2 (2)
	AAA module	1	1
	hsp60/GroEL chaperonin	1	1
	hsp20	2 (2)	0
	GARS (purine synthesis)	1	0

as well as alternative tranduction pathways for different effector systems; and for developmental specializations.

Distribution of Well-Studied Building Blocks

This table reflects the current state of building block classification; as new sequences become available and comparison methods improve, the percentage of proteins classified into families increases and new families are delineated. A biologially meaningful comparison can be made by counting the occurrence of particular modules in different organisms. Some ules are found in all organisms examined, whereas particular organisms may favor different modules for an analogous (e.g., regulation of transcription). In the largest family, the C_2H_2 Zn finger proteins, the module has diverse functions. vely, GARS is an example of a true ortholog, performing the identical enzymatic function in the different organisms shown. counts were obtained from OWL v.29.3 and family counts from ProClass v.1.1. Modules were tallied by searching OWL or 3L protein database, with exclusion of redundant entries from the final count.

	M. genitalium	*M. jannaschii*	*S. cerevisiae*	*A. thaliana*	*D. melanogaster*	*C. elegans*	*M. musculus*	*H. sapiens*
	470	1,734	6,297	1,189	1,566	11,274	7,161	11,060
	Complete	Complete	Complete	10	10	70	10	15
	9	5	25	51	50	28	63	57
teins Containing the Module)								
	0	0	96 (45)	10 (10)	200 (41)	365 (82)	717 (87)	1,323 (209)
	0	0	8 (8)	25 (25)	57 (55)	61 (61)	124 (118)	104 (101)
	0	0	52 (52)	0	0	0	0	0
	0	2 (2)	0	0	0	0	0	0
	0	2 (1)	2 (1)	4 (2)	4 (2)	4 (2)	2 (1)	2 (1)
	0	0	0	0	17 (17)	79 (79)	109 (109)	252 (252)
	1	0	114 (114)	125 (119)	157 (154)	294 (272)	331 (307)	562 (529)
	0	0	2 (2)	3 (2)	0	0	0	0
	0	0	0	0	2 (2)	2 (1)	32 (11)	127 (42)
	0	0	9 (6)	4 (4)	1	17 (11)	18 (10)	28 (18)
	0	0	6 (6)	4 (4)	12 (12)	16 (16)	12 (12)	11 (11)
	0	0	6 (3)	2 (1)	11 (7)	19 (13)	31 (31)	81 (46)
	1	0	9 (7)	1	3 (3)	6 (6)	14 (12)	15 (10)
	17 (16)	17 (16)	48 (30)	2 (1)	9 (6)	59 (38)	19 (12)	25 (16)
	2 (2)	2 (2)	26 (26)	2 (2)	7 (7)	17 (17)	8 (8)	11 (11)
	1	4 (3)	23 (20)	2 (1)	3 (3)	15 (13)	8 (7)	7 (7)
	1	1	9 (9)	3 (3)	2 (2)	6 (6)	9 (9)	8 (8)
	0	2 (2)	1	7 (7)	7 (7)	14 (14)	3 (3)	4 (4)
	0	1	1	1	1	1	1	1