




























Eubacteria

	<i>M. genitalium</i>		<i>B. subtilis</i>		<i>E. coli</i>	
Rank	Superfamily	#	Superfamily	#	Superfamily	#
1	 P-loop hydrolase	60	 P-loop hydrolase	173	 P-loop hydrolase	191
2	 SAM methyl-transferase	16	 Rossmann domain	165	 Rossmann domain	158
3	 Rossmann domain	13	 Phosphate-binding barrel	79	 Phosphate-binding barrel	64
4	Class I synthetase	12	 PLP-transferases	44	 PLP-transferases	38
5	Class II synthetase	11	 CheY-like domains	36	 CheY-like domains	36
6	Nucleic acid binding dom.	11	 SAM methyl-transferase	30	 Ferredoxins	35
Total ORFs		479		4268		4268
with common superfamilies		105 (22%)		465 (11%)		458 (11%)

Archaea

	<i>M. thermoautotrophicum</i>		<i>A. fulgidus</i>	
Rank	Superfamily	#	Superfamily	#
1	 P-loop hydrolyase	93	 P-loop hydrolyase	118
2	 Phosphate-binding barrel	54	 Rossmann domain	104
3	 Rossmann domains	53	 Phosphate-binding barrel	56
4	 Ferredoxins	48	 Ferredoxins	49
5	 SAM methyl-transferase	17	 SAM methyl-transferase	24
6	 PLP-transferases	15	 PLP-transferases	18
Total ORFs		1869		2409
with common superfamilies		252 (14%)		309 (13%)

Eukaryotes

	<i>S. cerevisiae</i>		<i>C. elegans</i>	
Rank	Superfamily	#	Superfamily	#
1	Δ P-loop hydrolyase	249	X Protein kinase	429
2	X Protein kinase	123	Δ P-loop hydrolyase	411
3	\otimes Rossmann domain	90	Ligand-binding NR dom.	254
4	RNA-binding domain	75	C-type lectin	253
5	= SAM methyl-transferase	63	alpha/beta hydrolase	180
6	Ribonuclease H-like	57	Ig superfamily	149
Total ORFs		6218		9,099
with common superfamilies		560 (9%)		1676 (8%)

Rough Layout for Table 3

Eubacteria

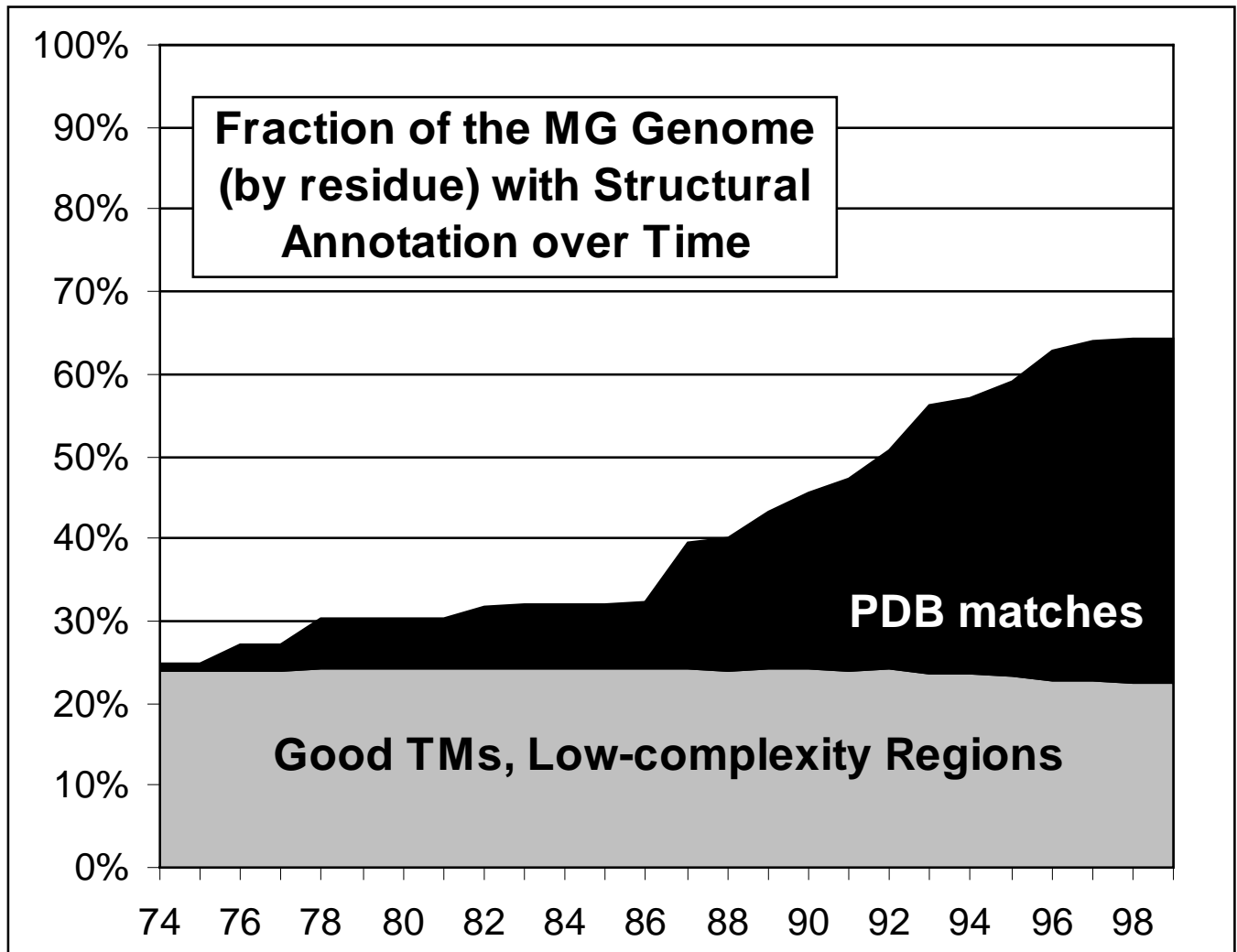
Rank	<i>M. genitalium</i>		<i>B. subtilis</i>		<i>E. coli</i>	
	Superfamily	#	Superfamily	#	Superfamily	#
1	△ P-loop hydrolase	60	△ P-loop hydrolase	173	△ P-loop hydrolase	191
2	= SAM methyl-transferase	16	⊗ Rossmann domain	165	⊗ Rossmann domain	158
3	⊗ Rossmann domain	13	● Phosphate-binding barrel	79	● Phosphate-binding barrel	64
4	◆ Class I synthetase	12	◆ PLP-transferases	44	◆ PLP-transferases	38
5	◆ Class II synthetase	11	* CheY-like domains	36	* CheY-like domains	36
6	◆ Nucleic acid binding dom.	11	= SAM methyl-transferase	30	◇ Ferredoxins	35
Total ORFs		479		4268		4268
with common superfamilies		105 (22%)		465 (11%)		458 (11%)

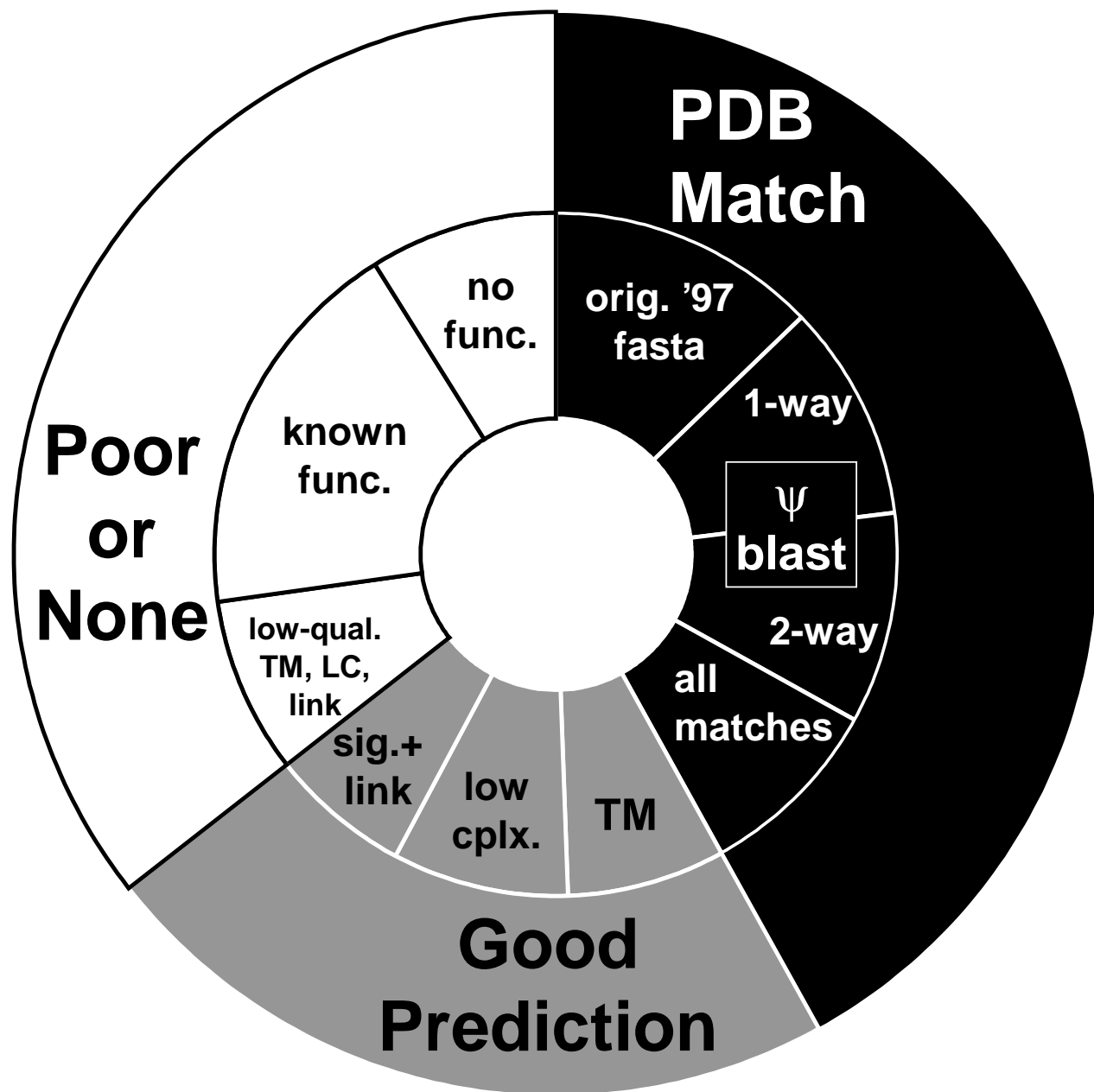
Archaea

Rank	<i>M. thermoautotrophicum</i>		<i>A. fulgidus</i>	
	Superfamily	#	Superfamily	#
1	△ P-loop hydrolyase	93	△ P-loop hydrolyase	118
2	● Phosphate-binding barrel	54	⊗ Rossmann domain	104
3	⊗ Rossmann domains	53	● Phosphate-binding barrel	56
4	◇ Ferredoxins	48	◇ Ferredoxins	49
5	= SAM methyl-transferase	17	= SAM methyl-transferase	24
6	◆ PLP-transferases	15	◆ PLP-transferases	18
Total ORFs		1869		2409
with common superfamilies		252 (14%)		309 (13%)

Eukaryotes

Rank	<i>S. cerevisiae</i>		<i>C. elegans</i>	
	Superfamily	#	Superfamily	#
1	△ P-loop hydrolyase	249	✕ Protein kinase	429
2	✕ Protein kinase	123	△ P-loop hydrolyase	411
3	⊗ Rossmann domain	90	⊗ Ligand-binding NR dom.	254
4	◆ RNA-binding domain	75	◆ C-type lectin	253
5	= SAM methyl-transferase	63	= alpha/beta hydrolase	180
6	◆ Ribonuclease H-like	57	◆ Ig superfamily	149
Total ORFs		6218		9,099
with common superfamilies		560 (9%)		1676 (8%)





Type of Structural Annotation	Number of ORFs with annotation	Additional frac.of tot. residues with annotation	Description of Methods Used to Determine Annotation
orig. '97 fasta	90	13%	This is the baseline structural annotation: MG regions matched to PDB domain structures in 1997 based on using FASTA [5] with a very conservative e-value threshold of .01 and an old database (scop 1.35) [8].
ψ-BLAST 1-way	161	10%	Additional regions matched to PDB structures (beyond above) based on running PSI-BLAST [19]. A more recent 1998 version of the PDB domains (scop 1.38 excluding coiled-coils and small Leu- and Cys-rich proteins) was run against MG embedded in NRDB (which had been masked in the default fashion by SEG [56]). These comparisons used 20 iterations, an inclusion threshold into the matrix of .0005, an overall match cutoff of .0001, and matches were continuously parsed from output.
ψ-BLAST 2-way	223	10%	Additional matches from running PSI-blast in two-way fashion plus pre-clustering the ORFs in MG [21]. By "two-way," we mean that the PDB was first run against MG embedded in NRDB and then unmatched regions of MG were cut out and run against the PDB embedded in NRDB. The pre-clustering was done with GEANFAMMER [24].
all matches	242	9%	Additional matches by considering the all the MG matches discussed in table 1 (i.e. various PSI-blast approaches, threading, etc [20,22,26,32]).
TM	79	7%	Surest annotation for TM-helices in integral membrane proteins. These were segments of at least 20 residues with an average GES hydrophobicity less than -1 kcal/mole [13, 39] in a protein that had at least one TM-segment with an average hydrophobicity less than -2 kcal/mole. (This adapted from Boyd & Beckwith's MaxH criteria [40].) Only about 7% of the residues are flagged as sure TM segments, but these occur in ~17% of the sequences.
low cplx.	65	8%	Very long low-complexity regions. These are thought not to fold into globular protein structures. They were identified with the SEG program using a trigger complexity K(1) of 3.4, an extension complexity K(2) of 3.75, and a window of length 45 [56]. In addition, the whole low-complexity region had to longer than 150 residues.
sig.+link	258	7%	Hydrophobic signal sequences and linking peptides. Signal sequences have the pattern of a charged residue within first seven followed by a stretch of 14 hydrophobic residues. Segments of sequence already accounted for thus far, i.e. PDB matches, low complexity, or TM-helices, are considered to be "characterized" regions. Short sequences (<80 residues) between characterized segments are considered to be linkers.
low qual. TM, LC, link	42	9%	This category consists of much lower quality structural annotation for TM-helices and low-complexity regions. That is, low-complexity region by the same criteria as discussed above but shorter than 150 amino acids, and TM-helices with an average hydrophobicity less than -1 kcal/mole but are in proteins that do not meet the MaxH criteria.
known func.	131	18%	These regions have no other structural annotation but occur in protein given functional annotation by Mushegian & Koonin [55] or TIGR [4] and thus probably fold into globular structures.
no func.	70	9%	Region has no structural annotation and occurs in a protein that is given no functional annotation by TIGR or Mushegian & Koonin (as of Jan-99).