Pacific Symposium on Biocomputing 7:127-138 (2002)

STRUCTURE-BASED COMPARISON OF FOUR EUKARYOTIC GENOMES

MELISSA CLINE, GUOYING LIU, ANN E. LORAINE, RONALD SHIGETA,
JILL CHENG, GANGWU MEI, DAVID KULP, and MICHAEL A. SIANI-ROSE
Affymetrix, Inc. 6550 Vallejo Street,
Emeryville, CA 94608 USA
E-mail: melissa_cline@affymetrix.com

The field of comparative genomics allows us to elucidate the molecular mechanisms necessary for the machinery of an organism by contrasting its genome against those of other organisms. In this paper, we contrast the genome of *homo sapiens* against *C. Elegans, Drosophila melanogaster, and S. cerevisiae* to gain insights on what structural domains are present in each organism. Previous work has assessed this using sequence-based homology recognition systems such as Pfam [1] and Interpro [2]. Here, we pursue a structure-based assessment, analyzing genomes according to domains in the SCOP structural domain dictionary. Compared to other eukaryotic genomes, we observe additional domains in the human genome relating to signal transduction, immune response, transport, and certain enzymes. Compared to the metazoan genomes, the yeast genome shows an absence of domains relating to immune response, cell-cell interactions, and cell signaling.

1 Introduction

To date, there are hundreds of completed genomes for prokaryotes, but only five for eukaryotes: *homo sapiens, Caenorhabditis elegans, Drosophila melanogaster, Saccharomyces cerevisiae,* and *Arabidopsis thaliana.* Eukaryotes exhibit more genomic complexity. Their genomes are larger, their proteins contain a wider variety of domains, and the domains appear in a greater number of combinations [3]. Yet compared to other eukaryotes, the human genome exhibits even more complexity. The human proteome contains a greater number of domains and domain combinations than other eukaryotic proteomes [2]. Compared to other sequenced eukaryotes, the human genome yields more immune response proteins, epithelial proteins, and olfactory [2], plus more proteins related to neural development and function, signaling, homeostasis, and apoptosis [1].

While some amount has been written about the eukaryotic genomes from a functional standpoint[1, 2], little has been published in terms of comparative structural genomics. Historically, structural analysis has been used largely as an intermediate step towards the goal of functional analysis, although this has its limitations[4, 5]. A structure-based analysis is worthwhile in its own right. First, while structural similarity is a clearly defined concept, functional similarity is more ambiguous. The many definitions of functional similarity include common domains, common EC numbers, similar keywords in Medline abstracts, and similar binding sites; whether or not two proteins are considered functionally similar depends on which definition of functional similarity is used. Second, structural classification schemes capture different information than functional classification schemes [6]. A comparison of SCOP[7] and Pfam[8] domains found that 70% of the Pfam families have corresponding entries in SCOP, while 57% of the SCOP families have corresponding entries in Pfam[9]. Third, structural classification schemes can be organized hierarchically, yielding a convenient way to summarize the content of a genome. Fourth, SCOP has been used in previous studies to examine genomes [5, 10-13], and the total number of folds in SCOP have been carefully examined [14, 15]. Thus, functional and structural genomic analyses are good complements.

To perform a structural genomics analysis, we applied the GRAPA method [16]. GRAPA features a library of HMMs based around the SCOP hierarchy, with each HMM optimized for family recognition, a contrast to other methods optimized for superfamily recognition [17, Karplus, 1998 #16]. In GRAPA, a protein is assigned to a SCOP family by comparing the distance scores of each candidate for each of the HMMs within a SCOP superfamily. We used this library to analyze four eukaryotic genomes: *homo sapiens, C. elegans, drosophila melanogaster,* and *S. Cerevisiae.* To put our results into perspective with prior work, we applied the same genomes to the Pfam library. Using these results, we contrasted the human genome to other eukaryotic genomes in both a functional and structural perspective.

2 Methods

2.1 Genome Gene Sets

A set of protein sequences covering the Golden Path of the human genome (October 7, 2000 freeze http://genome.ucsc.edu/) was generated by the Genie [18] programs suite [Kulp, D. & Wheeler, R., published in http://genome.ucsc.edu/], with the repeat regions masked out. The data set consists of three sets of amino acid sequences: (1) proteins from Genbank whose associated mRNA sequences could be mapped to genomic sequence (2) proteins predicted by *AltGenie*, an enhanced Genie program which predicts alternatively splice transcripts using merged mRNA/EST-to-genomic alignments and (3) proteins predicted by *StatGenie*, a purely *ab initio* gene-finder. These sequences are non-redundant; none of the included genes overlap the same genomic region. In cases where there were many genes overlapping the same region, the one with the longest CDS (translation) was kept [Williams, A., unpublished data]. This set, known as annot10, contains 59,378 protein sequences.

The non-redundant complete proteome set of SWISS-PROT plus TrEMBL entries for *Drosophila melanogaster* (13844 entries), *Caenorhabditis elegans* (18870 entries), and *Saccharomyces cerevisiae* (6186 entries) were obtained on June 15, 2001 from the EBI proteome analysis site (http://www.ebi.ac.uk/proteome/).

2.2 Model Building and Search Method

We searched for domains in the target genome according to GRAPA, a battery of Hidden Markov Models (HMMs) generated for each member of the SCOP hierarchy. GRAPA characterizes each SCOP protein domain found in ASTRAL[19], a service that allows the user to select proteins from SCOP according to various criteria. We selected all non-redundant proteins from SCOP version 1.53, yielding 4369 entries. For each entry, a hidden Markov model (HMM) was built using the Target99 protocol [20] with the Sequence Alignment and Modeling system (SAM 3.0) system[21]. Multiple species were included to capture characteristics of both mammalian and non-mammalian proteins.

Each gene was scored against each SCOP family, yielding an e-value. In practice, the e-values generated by a model are dependent in part on that model, with shorter models yielding higher e-values. This is consistent with the definition of the e-value: the number of equivalent or better scores that might arise by chance from non-homologous sequences, given a database of the same size. If a model is shorter, its best hits will be shorter; a shorter hit is more likely to occur by random chance than a longer one.

Because the e-value interpretation is dependent on the HMM, there is no single E-value that can be applied to all models. Instead, for each HMM, the DISTSIEVE program examines a model's set of e-values and determines a reasonable e-value cutoff by curve analysis. In general, the hits to an HMM will include a small number of hits with low e-values, corresponding to the training sequences and their homologs, followed by a large increase in the number of hits as the threshold rises to include the false positives. DISTSIEVE examines the hits to each model and identifies some e-value threshold beyond which the number of hits increases rapidly. The hits selected by DISTSIEVE are then assigned to whatever model in each superfamily they score against best. The performance of this method is comparable to PSI-BLAST, with a family recognition specificity of approximately 95% and sensitivity of approximately 75%.

For comparison with previous work, we searched Pfam with the same sequences. We applied the genomes to Pfam 6.2, recording all hits, which exceeded each model's, gathering threshold. The gathering threshold is defined by the Pfam authors as the scoring threshold above which they would admit a new sequence to the Pfam alignment. There is one gathering threshold per model, and it is set manually.

To establish a correspondence between the SCOP and Pfam domain definitions, we followed the method applied in previous work [9] and scored the SCOP domain sequences against all Pfam models. When the score of a SCOP sequence exceeded the gathering threshold for a Pfam model, we noted the hit as a potential correspondence. We then examined the potential correspondences by hand for the domains emphasized in this paper.

3 Results and Discussion

We have studied the types of structural domains found within four genomes. Table 1 lists the number of different SCOP domain families found within each genome. As expected, we see in Table 1 that the human genome is the most complex, with 442 families of structural domains represented. In rough terms, this approximates the number of different functions performed by the genes in the organism's genome.

Table 1. Number of different domain families found in each organism.

Organism	Number of Domain Families
Human	442
Fly	389
Worm	378
Yeast	245

Table 2 lists the twenty most frequent SCOP domain families in the human genome, and lists the number of occurrences of each domain within each organism. The corresponding Pfam domain and rank in the human genome is provided for comparison; in general, these numbers are equivalent to those previously published [22, Lander, 2001 #18, Venter, 2001 #32]. Mostly, the top-ranked domains are similar for SCOP and Pfam. One contrast concerns Leucine Rich Repeats, LRRs. These short sequence motifs appear in many different types of proteins and many different SCOP domains, including Internalin B, B LLR domain, U2A'-like, and Rna1p. Pfam, in contrast, includes a separate LLR model. Another contrast concerns proteases. Eukaryotic and prokaryotic proteases, which are among the more common SCOP domains, have no direct equivalent in Pfam. The closest Pfam model, the trypsin domain, appears to be more specific and would likely not rank in the top twenty. Immunoglobulin domains appear as a top hit under Pfam (rank = 4), while SCOP finds many such hits via the V set domains (antibody variable domain-like) (rank=19). Closer examination revealed that the SCOP model is more specific.

Table 2. Number of human proteins (H), C. elegans (W), Drosophila melanogaster (F), and S. cerevisiae (Y) sorted by most frequently occurring SCOP v 1.53 families in humans. The rank of the corresponding Pfam model (if any) is shown for comparison.

	SCOP Family Id	H	F	W	Y	SCOP family	Pfam rank equivalent
1	4.130.1.1	349	220	422	113	Serine/threonine kinases	pkinase (3)
2	7.37.1.1	296	185	56	27	Classic zinc finger, C2H2	zf-C2H2 (1)
3	4.130.1.2	246	34	399	114	Tyrosine kinase	pkinase (3)
4	3.32.1.8	184	128	147	72	G proteins	ras (20)
5	1.111.2.1	135	84	92	18	Ankyrin repeat (SH3-domain superfamily)	ank (5)
6	2.64.3.1	115	21	11	15	Trp-Asp repeat (WD-repeat)	WD40 (8)
7	3.9.2.1	99	119	35	8	Internalin B LRR domain	LRR (9)

8	3.9.2.3	97	108	46	9	U2A'-like Leucine Rich Repeat Fold, RNA recog.	LRR (9)
9	3.32.1.13	90	58	69	52	Extended AAA-ATPase domain (DNA helicases bacterial/yeast)	helicase_C (26) and DEAD (27)
10	3.9.1.2	86	103	15	12	Rna1p (in Leucine Rich Fold)	LRR (9)
11	2.44.1.2	82	208	12	0	Eukaryotic proteases	-
12	2.44.1.1	78	200	3	1	Prokaryotic proteases	-
13	1.4.1.1	74	96	78	7	Homeodomain	homeobox (14)
14	4.37.1.1	73	74	98	3	BTB/POZ domain (zinc finger)	BTB (24)
15	1.23.1.1	65	7	51	9	Nucleosome core histones	histone (29)
16	3.9.1.1	64	57	11	10	Ribonuclease inhibitor (LRR fold)	LRR (9)
17	4.82.1.1	63	39	65	0	SH2 domain	SH2 (25)
18	3.9.2.2	63	104	59	8	Rab geranylgeranyltransferase a-subunit, N-terminal (C2 domain-like Fold)	C2 (22)
19	2.1.1.1	61	70	26	0	V set domains (Ab variable domain-like) Ig superfamily	Ig (4)
20	3.32.1.9	61	41	36	11	Motor proteins (nucleoside triphosphate hydrolase family)	myosin head motor domain (52)

The Zinc finger C3HC4 type RING domain (rank = 13) in the Pfam top twenty, while the corresponding SCOP RING model has a rank of ninety. Further examination would be required to determine whether these Pfam hits appeared in the SCOP C2H2 (rank = 2) or BTB/POZ zinc finger (rank = 14) domains. In the case for the Pfam EF-HAND domain (rank = 15), SCOP has broken the EF-hand superfamily into seven families, which would yield fewer hits per family and therefore make them less likely to appear in the top twenty models.

One implication of the data in Table 2 is the prevalence of signaling proteins in the human genome: kinases, proteases, G proteins, and so forth. This reflects the importance and variety of signaling mechanisms within higher-order organisms such as humans. Much of this emphasis on signaling proteins is also evidenced in the worm genome; much less is present in the genomes of fly and yeast.

The top twenty Pfam entries (data not shown) with no corresponding SCOP entry, correspond to transmembrane proteins (7tm_1) and other entries for which there are no solved 3D structures.

Table 3. Top structural domain families common to Metazoan genomes and absent from yeast.

	SCOP	H	F	W	Y	SCOP family

	family designation					
1	2.44.1.2	82	208	12	0	Eukaryotic proteases
2	4.82.1.1	63	39	65	0	SH2 domain
3	2.1.1.1	61	70	26	0	V set domains (antibody variable domain-like)
4	2.1.2.1	57	40	25	0	Fibronectin type III
5	2.1.1.4	50	117	56	0	I set domains
6	2.1.6.1	48	15	15	0	Cadherin
7	4.154.1.1	41	34	224	0	C-type lectin domain
8	7.3.10.1	39	13	13	0	EGF-type module
9	1.116.1.1	36	18	185	0	Nuclear receptor ligand-binding domain
10	3.57.1.1	34	3	30	0	Integrin A (or I) domain
11	7.39.1.2	34	22	246	0	Nuclear receptor
12	2.2.5.1	32	18	11	0	p53-like transcription factors
13	2.34.1.2	31	51	41	0	Interleukin 16
14	7.17.1.2	29	7	4	0	Transforming growth factor (TGF)-beta
15	4.37.1.2	29	12	46	0	Tetramerization domain of Potassium Channels

Table 3 lists the structural domain found most frequently in human, fly, and worm, and not found in yeast. Not surprisingly, eukaryotic proteases figure prominently. This family of proteins is exclusive to eukaryotes and includes trypsin, chymotrypsin, neuropsin, collagenase, thrombin, carboxypeptidase, elastase, enteropeptidase, heparin binding protein, beta-tryptase, chathepsin G, coagulation factors VIIa and Xa, kallikrein A, tonin, Nerve Growth Factor, Factor D, plasminogen activator, activated protein c, myeloblastin, and plasminogen.

Many of the structural domains in Table 3 are involved in immune responses. These include proteases, SH2 domains, V-set domains, I-set domains, and nuclear receptors. Both V-set and I-set domains belong to the Immunoglobulin V set domain (antibody variable domain-like) superfamily. Fibronectin Type III family contains many immune system receptors. C-type lectin domains are found in Natural Killer cell receptors and other immune system cell-recognition cell-surface receptors. While these motifs would not have the same immunological function in Fly and Worm, as in humans, the structural elements are clearly present in these genomes.

Proteins specifically associated with multicellular organisms include: cadherin (cell adhesion protein), and integrin-A (or I) domains are involved in cell-cell interaction.

Other structural domains in Table 3 are involved in cell signaling. Cell signaling involves both extracellular signaling proteins, cell surface receptors, and signaling pathways which transmit the signal within the cell. EGF-type modules are found in epidermal growth factor and many other growth factor and hormone extracellular signaling proteins. Transforming growth factor-beta (TGF-beta) and the cytokine Interleukin 16 are intercellular signaling proteins. Cell surface receptors include I-set domains found in Natural killer cell receptors. Intracellular signaling cascades are often regulated by SH2 (Src homology 2) domains; they interact with high affinity to phosphotyrosine-containing target peptides in a sequence specific and phosphorylation-dependent manner. Furthermore, intracellular signaling proteins that affect transcription include the nuclear receptors. In addition P53 transcription factors are involved in tumor suppression, a feature one would more likely expect in multicellular organisms.

Of particular interest are the high number of I set domains found in Fly (117 verses 50 in human). Also, the high number of C-type lectin in Worm (224 verses 41 in human) and nuclear receptors (246 verses 34 in human) suggests evolutionary branching.

Table 4. Top ten SCOP superfamilies unique to human with relevant Gene Ontology function and process annotations. The column labeled H contains the number of human genes in the indicated SCOP superfamily.

SCOP superfamily Id	H	SCOP superfamily title	GO Molecular Function	GO Molecular Process
11.27.1	40	4-helical cytokines	Signal transducer; ligand; growth factor	Cell growth & maintenance; stress response
24.18.1	30	MHC antigen-recognition domain	Signal transducer;transmembrane receptor	cell growth & maintenance; stress response
32.11.1	26	Lipase/lipooxygenase domain	Enzyme; hydrolase acting ester bonds	cell growth & maintenance; protein metabolism and modification
44.9.1	16	Interleukin 8-like chemokines	Enzyme; transferase for phosphorus groups	cell growth & maintenance; response to abiotic stimulus
54.72.1	12	Bactericidal permeability-increasing protein	Defense/immunity protein; antimicrobial response protein	cell growth & maintenance; response to external stimulus
67.43.1	11	B-box zinc-binding domain	Enzyme; transferase for phosphorus groups	cell growth & maintenance; developmental processes; metabolism
74.16.1	10	Cystatin/monellin	Enzyme inhibitor; proteinase inhibitor	cell growth & maintenance; response to external stimulus
87.24.1	10	TNF receptor-like	Signal transducer; transmembrane receptor	cell growth & maintenance; response to external stimulus

97.31.1	10	GLA-domain	Enzyme; hydrolase	cell communication; signal transduction
107.23.1	9	TB module/8-cys domain	Structural protein; ligand binding or carrier	cell growth & maintenance; response to external stimulus

To see what structures are unique to higher-order organisms such as humans, Table 4 lists the top ten superfamily domains that are unique to the human genome. Most of these domains are involved in immune responses. Mapping of the SCOP families unique to humans back to LocusLink[23] and then into the Gene Ontology[24] (GO) graphs, allows one to rapidly clarify the role these proteins play. In Table 4, one can readily see the relationship between the SCOP structure superfamily name and the Gene Ontology Molecular Function and Process categories. The proteins unique to humans are involved in signal transduction (both ligand and receptor), enzymes (including various growth factors and cytokines, oxidoreductases, transferases, and hydrolases), miscellaneous defense and immunity proteins, transporter proteins, structural proteins, and ligand binding or carrier proteins.

The superfamily with the most hits by far is SCOP 1.27.1, 4-helix bundle cytokines, with 40 genes; all appear to be unique to human. This structural class, including long-chain cytokines, short-chain cytokines, and Inteferons/interleukin-10, is responsible in humans for mediating immune response across different organs and tissues and is responsible for much our highly evolved immune system. The second largest set, SCOP 4.18.1 superfamily, consists of the MHC antigen recognition domain, which are involved in training immune system cells to recognize foreign proteins. The fourth largest set, SCOP 4.9.1 superfamily consists of the Interleukin-8-like chemokines, another set of signaling proteins involved in lymphocyte trafficking. Interestingly, the incorporation of GO into this process, lets us identify more genes as potential cytokines by the highly nested functional notation of the GO graph. In addition to the explicitly cytokine superfamilies (1.27.1 and 4.9.1), other domains with implicit cytokine activity can be found: (1) SCOP superfamily 7.25.1, under the heparin-binding domain from vascular endothelial growth factor; (2) SCOP superfamily 4.36.3 alpha/beta-hammerhead pyrimidine nucleoside phosphorylase C-terminal domain; (3) SCOP superfamily 3.21.1 pyrimidine nucleoside phosphorylase central domain; and (4) SCOP superfamily 1.48.2 pyrimidine nucleoside phosphorylase N-terminal domain (methionine synthase domain-like). Interestingly these last three distinct domains, all part of pyrimidine nucleoside phosphorylase, are unique to human, indicating a whole human protein consisting of three separate unique structural domains. In human beings, thymidine phosphorylase (TP) performs metabolic functions in degradation of various drug compounds as well as being overexpressed in many tumor types and linked to angiogenesis. While the exact role that TP plays is not biochemically characterized, it seems that humans have adapted TP to a signaling role. The enzyme's absence in lower metazoans implies that TP may have been the result of a lateral gene transfer from a bacterium.

Table 5. Genes with disproportionate numbers between genomes. These figures represent genes that are annotated with greater than or equal to 20% sequence Identity to the SCOP seed sequence (after being annotated by GRAPA HMM scoring. This is intended to ensure that all hits are real.

SCOP Id	Human	Fly	Worm	Yeast	SCOP family
7.39.1.2	34	22	210	0	Nuclear receptors
2.44.1.2	71	163	3	0	Eukaryotic proteases
2.44.1.1	60	156	2	0	Prokaryotic proteases
1.23.1.1	63	5	49	7	Nucleosome core histones
1.23.1.2	16	6	30	6	Archaeal histone
1.4.5.12	9	1	8	0	Histone H1/H5
4.37.1.2	26	10	23	0	Tetramerization domain of potassium channels
4.145.1.3	6	13	32	4	Protein serine/threonine phosphatase
4.154.1.1	21	4	22	0	C-type lectin domain
1.111.6.1	3	2	2	1	Protein prenylyltransferase
3.9.2.2	28	39	13	1	Rab geranylgeranyltransferase alpha-subunit, N-terminal domain

By examining SCOP families disproportionately represented among the three metazoans (see Table 5), we can identify some points where other families of proteins might have fulfilled some of the signaling functions observed only in mammals. Examples of this include cytokines and immune signaling pathways. The large number of nuclear receptors (7.39.1.2) in worm suggests that the worm might rely on hormonal small molecules for development rather than intracellular signaling proteins. This disproportionate number of nuclear receptors has been observed before[2]. Another example concerns the expansion of histone components (SCOP families 1.23.1.1, 1.4.5.12, and 1.23.1.3) *in C. elegans.* This suggests that perhaps gene regulation in *C. elegans* might be more reliant in histone packing and modification for cell signaling. The appearance of K-channel associated domains (4.37.1.2) and signaling/transport protein phosphatases (4.145.1.3) also show diversity in signaling/transport pathways in *C. elegans*

The worm also has a plethora of C-type lectins (4.154.1.1), which are responsible for mediating intracellular contact information through surface carbohydrates.

Expansions of protease families (2.44.1.2) and families involved in protein prenylation (1.111.6.1 and 3.9.2.2) in *Drosophila* suggest an emphasis on regulated, post-translational protein modifications in fly. Protein prenyl transferases attach hydrophobic prenyl groups to nuclear lamins as well as signaling molecules including ras superfamily members and the gamma subunits of trimeric G-proteins, all of which require this modification to attach to membranes and interact with effector molecules. Regulated prenylation has been demonstrated for farnesyl transferase in human [25] but has not been demonstrated for Rab prenyl transferase, which prenylates Rab proteins, a subcategory within the Ras superfamily and which help regulate discrete steps in the secretory pathway. Rab prenyl transferase is present in yeast, fly, worm, and human and requires a regulatory subunit (Rab Escort Protein) to bind and present the protein substrate to the α/β catalytic subunits of the enzyme [26, 27]. Interestingly, both fly and human possess a large number of proteins (fly: 104, human: 63) recognized by the single SCOP-HMM model trained on a structural motif present in the N-terminal region of the Rab prenyl transferase alpha subunit (3.9.2.2), a section of the protein proposed to interact with REP [28]. Expansion of the protease families in fly has already been discussed [2], but over-representation of proteins in fly and human exhibiting structural homology to protein prenyl transferases motifs has not been reported until now.

4 Conclusions

The SCOP structural domain hierarchy allows us to characterize diverse genomes in a complementary manner to previous work involving Pfam and Interpro. The two methods provide a consistent "view" of the most highly represented genes in the human, fly, worm, and yeast genomes.

Structure-based genome comparison, as provided by SCOP domain families, allows us to track the appearance or elimination of structural domains. Coupled with a system such as Gene Ontology, this method allows us to find related genes related via common structure that diverge in function or in expression pathway. We find that metazoans diverge from the yeast genome by a set of structure-based domains consistent with previous observations. These novel domains correspond to the well-characterized eukaryotic proteases, and processes and function associated with immune response-related, cell-cell interaction, and cell signaling pathways. Interestingly, some families of proteins which vary widely between genomes (e.g., nuclear receptors) appear much more within one genome than within another. These allow us to track evolutionary biases due to different uses of the same basic structural features.

To date, few of the genomes sequenced are eukaryotic. However, as shown here and in previous work [2, 3, 12], eukaryotic genomes exhibit a different composition of domains than prokaryotic genomes. Thus, if genomic databases are largely prokaryotic, researchers should bear in mind that such databases will have a decreased representation of certain classes of proteins important in eukaryotes and absent in prokaryotes, such as those involved in intercellular signaling, immune response, and cell-cell interactions.

5 References

1. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.
2. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
3. Apic, G., J. Gough, and S.A. Teichmann, *Domain combinations in archaeal, eubacterial and eukaryotic proteomes.* J Mol Biol, 2001. **310**(2): p. 311-25.
4. Devos, D. and A. Valencia, *Practical limits of function prediction.* Proteins, 2000. **41**(1): p. 98-107.
5. Wilson, C.A., J. Kreychman, and M. Gerstein, *Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores.* J Mol Biol, 2000. **297**(1): p. 233-49.
6. Gerstein, M. and R. Jansen, *The current excitement in bioinformatics-analysis of whole-genome expression data: how does it relate to protein structure and function?* Curr Opin Struct Biol, 2000. **10**(5): p. 574-84.
7. Murzin, A.G., et al., *SCOP: a structural classification of proteins database for the investigation of sequences and structures.* J Mol Biol, 1995. **247**(4): p. 536-40.
8. Bateman, A., et al., *The Pfam protein families database.* Nucleic Acids Res, 2000. **28**(1): p. 263-6.
9. Elofsson, A. and E.L. Sonnhammer, *A comparison of sequence and structure protein domain families as a basis for structural genomics.* Bioinformatics, 1999. **15**(6): p. 480-500.
10. Teichmann, S.A., C. Chothia, and M. Gerstein, *Advances in structural genomics.* Curr Opin Struct Biol, 1999. **9**(3): p. 390-9.
11. Gerstein, M. and H. Hegyi, *Comparing genomes in terms of protein structure: surveys of a finite parts list.* FEMS Microbiol Rev, 1998. **22**(4): p. 277-304.
12. Wolf, Y.I., et al., *Distribution of protein folds in the three superkingdoms of life.* Genome Res, 1999. **9**(1): p. 17-26.
13. Wolf, Y.I., N.V. Grishin, and E.V. Koonin, *Estimating the number of protein folds and families from complete genome data.* J Mol Biol, 2000. **299**(4): p. 897-905.
14. Govindarajan, S., R. Recabarren, and R.A. Goldstein, *Estimating the total number of protein folds.* Proteins, 1999. **35**(4): p. 408-14.
15. Zhang, C. and C. DeLisi, *Estimating the number of protein folds.* J Mol Biol, 1998. **284**(5): p. 1301-5.
16. Shigeta, R., et al., *Generalized Rapid Automated Protein Analysis (GRAPA): annotating the human genome based on SCOP domain-derived hidden Markov models.* submitted, 2001.

17. Gough, J., et al. *Optimal Hidden Markov Models for All Sequences of Known Structure*. in *Currents in Computational Molecular Biology 2000*. 2000.
18. Reese, M.G., et al., *Genie--gene finding in Drosophila melanogaster.* Genome Res, 2000. **10**(4): p. 529-38.
19. Brenner, S.E., P. Koehl, and M. Levitt, *The ASTRAL compendium for protein structure and sequence analysis.* Nucleic Acids Res, 2000. **28**(1): p. 254-6.
20. Karplus, K., C. Barrett, and R. Hughey, *Hidden Markov models for detecting remote protein homologies.* Bioinformatics, 1998. **14**(10): p. 846-56.
21. Hughey, R. and A. Krogh, *Hidden Markov models for sequence analysis: extension and analysis of the basic method.* Comput Appl Biosci, 1996. **12**(2): p. 95-107.
22. Rubin, G.M., et al., *Comparative genomics of the eukaryotes.* Science, 2000. **287**(5461): p. 2204-15.
23. Pruitt, K.D. and D.R. Maglott, *RefSeq and LocusLink: NCBI gene-centered resources.* Nucleic Acids Res, 2001. **29**(1): p. 137-40.
24. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
25. Goalstone, M.L., et al., *Insulin signals to prenyltransferases via the Shc branch of intracellular signaling.* J Biol Chem, 2001. **276**(16): p. 12805-12.
26. Armstrong, S.A., et al., *cDNA cloning and expression of the alpha and beta subunits of rat Rab geranylgeranyl transferase.* J Biol Chem, 1993. **268**(16): p. 12221-9.
27. Andres, D.A., et al., *cDNA cloning of component A of Rab geranylgeranyl transferase and demonstration of its role as a Rab escort protein.* Cell, 1993. **73**(6): p. 1091-9.
28. Zhang, H., M.C. Seabra, and J. Deisenhofer, *Crystal structure of Rab geranylgeranyltransferase at 2.0 A resolution.* Structure Fold Des, 2000. **8**(3): p. 241-51.