IUBMB *Life*, 56(6): 317–331, June 2004

Taylor & Francis
healthsciences

Critical Review

Structural and Functional Characterization of Gene Products Encoded in the Human Genome by Homology Detection

S. B. Pandit, S. Balaji and N. Srinivasan
Molecular Biophysics Unit, Indian Institute of Science, Bangalore-560012, India

Summary

Availability of the human genome data has enabled the exploration of a huge amount of biological information encoded in it. There are extensive ongoing experimental efforts to understand the biological functions of the gene products encoded in the human genome. However, computational analysis can aid immensely in the interpretation of biological function by associating known functional/structural domains to the human proteins. In this article we have discussed the implications of such associations. The association of structural domains to human proteins could help in prioritizing the targets for structure determination in the structural genomics initiatives. The protein kinase family is one of the most frequently occurring protein domain families in the human proteome while P-loop hydrolase, which comprises many GTPases and ATPases, is a highly represented superfamily. Using the superfamily relationships between families of unknown and known structures we could increase structural information content of the human genome by about 5%. We could also make new associations of domain families to 33 human proteins that are potentially linked to genetically inherited diseases.

IUBMB *Life*, 56: 317–331, 2004

Keywords Genome analysis; homology detection; protein domain folds; protein evolution; sequence analysis.

INTRODUCTION

An important event in the history of humankind has been made at the turn of the 21st century in the form of the draft version of human genome data (*1, 2*). This monumental effort provides us with an opportunity to better understand the various biological processes and possible cognition. The functional characterization of the gene products encoded in the human genome could provide insights into such complex biological process. There are various experimental endeavors to understand the functions of gene products encoded in the human genome. In addition to these experimental efforts, computational analyses of human proteins could form an important step in the functional inference of the genome data. The computational approaches for the prediction of functional features of proteins encoded in genomes relies on establishing relationships to homologues that are experimentally studied. There have been several attempts, using various sophisticated homology search tools, to assign functions to gene products encoded in various proteomes (see for example references *3–8*). Such functional predictions could be used as a guiding tool in order to direct the relatively time consuming, more difficult and expensive experimental methods for exploring protein functions. Furthermore, functional inferences of human proteins that are implicated in diseases could provide valuable insights on the molecular basis of human diseases. Such an understanding could aid identification of effective drug targets and rational design of lead compounds to combat the diseases.

The most commonly used computational approach for genome-wide association of functions to proteins is by identification of well-characterized homologues using sequence-based search procedures such as BLAST (*9*) and FASTA (*10*). But, pairwise sequence alignment based search procedures are unlikely to be able to identify related proteins with low sequence similarity. However, these distantly related proteins could often be identified with the use of three-dimensional (3-D) structural information (*11*) as the structure is conserved better than sequence during evolution (*12, 13*) Thus, use of structural information could potentially enhance the functional assignments (*14–17*). Moreover, structure prediction with relevant biochemical motifs can provide more detailed functional insights than sequence comparisons alone (*18–20*). The search methods for such an analysis could be improved by the use of multiple sequence alignment of the homologues in a family, which can indicate structurally/functionally important positions. The information in these multiple sequence alignments can be converted into Position Specific Scoring Matrices (PSSM) usually referred as profiles (*21*) or into a probabilistic model called the Hidden Markov Model (HMM) (*22*). The use of profile-based search methods is known to improve sensitivity of detection of remotely related homologues (*23–28*). Hence, use of structure and

Address correspondence to: N. Srinivasan, Molecular Biophysics Unit, Indian Institute of Science, Bangalore-560012, India. Tel: + 91-80-2293 2837. Fax: + 91-80-2360 0535. E-mail: ns@mbu.iisc.ernet.in

ISSN 1521-6543 print/ISSN 1521-6551 online © 2004 IUBMB
DOI: 10.1080/15216540400006105

profile-based method would enable detection of remote homologues and thus enrich the functional assignments.

Some of the commonly used profile-based search methods include PSI-BLAST (*25*), IMPALA (*29*), RPS-BLAST and Hidden Markov Model (HMMER2)-based (*30*) procedures. These methods have been shown to detect remote and subtle similarities (*31, 32*) between proteins that were previously possible only by structure comparison procedures, which obviously demand the knowledge of 3-D structures. Although the profile-based annotation methods are among the widely used procedures to detect remote similarities, there are other procedures such as GenThreader (*33*) and environment-based profiles (*34*), which are fold recognition methods for assigning the structural domains to amino acid sequences. Furthermore, the comparison of proteins across various genomes could also aid in enhancement of the assignment of functions to the proteins (*34–41*). The comparative genomics methods are based on the functional characterization of the gene products by detecting the orthologous proteins in the closely related organisms where the experimental functions of the proteins have been proposed. However, the effectiveness of this approach is dependent on at least two factors:

(1) Ability to identify homologues of a given protein in other organisms.
(2) The extent of divergence of amino acid sequences of the homologues across the organisms and its implication on the similarity of functions between the two proteins.

In order to understand the biological function of the human proteins we have associated functional or structural domains using sensitive profile-matching procedures. Association of functional domain would provide clues to biochemical role of the protein. The structural domain association could provide enhanced abilities to assign function and also provide the molecular basis of action of proteins. Furthermore, we have enhanced the structural information content of human genome by relating families apparently with unknown structures to known structural families, as in the SUPFAM database, which was developed by us earlier (*26, 27*). The functional domain information was considered from Pfam database (http://www.sanger.ac.uk/Software/pfam) (*42*) and structural domain from PALI (http://pauling.mbu.iisc.ernet.in/~pali) (*43, 44*) databases. PALI contains structure-based sequence alignment and phylogeny of proteins in the families derived from the SCOP database (*45*), which is a hierarchical structural classification database. The profiles for Pfam and PALI families have been generated as described by Pandit et al (*26*). These profiles were searched using RPS-BLAST. We have also used HMM-based search procedure against the HMM libraries of Pfam to associate functional domain to human proteins. Profiles of transmembrane sequences have been associated with human proteins in order to predict membrane localisation of the proteins.

Using sensitive profile-matching procedures, we could make a comprehensive compilation of functional/structural domains to gene protein encoded in human genome. Similar attempts have been made in the past by Muller et al. (*38*). They have used PSI-BLAST for much of their analysis, apart from IMPALA, to assign structural/functional domains. In their approach PSI-BLAST has been scanned against the non-redundant database of protein sequences augmented with SCOP domain sequences. In the present review we will discuss the current status of large-scale function association, using various computational procedures, of various gene products encoded in the human genome. Furthermore, human proteins potentially involved in diseases have been specifically analyzed by associating functional/structural domain to these protein sequences.

OVERVIEW OF DATA SET USED AND METHODOLOGY

The amino acid sequences of the Open Reading Frames (ORFs) that correspond to the gene products encoded in human genome have been obtained from the ENSEMBL database (*46*) (Release 22.34d.1, http://www.ensembl.org). The total number of gene products predicted in this release is 29,031.

The OMIM database (*47*) is a comprehensive collection of genes and genetic disorders in humans. In our analysis the protein sequences corresponding to the entries in the OMIM database have been derived from the SWISSPROT database (*48*). It is also possible to obtain details of genes involved in genetic disorders through the genelink table provided at ENSEMBL database (*46*). The genelink table indicates the association of human proteins to OMIM identifiers. We were able to associate 6257 unique protein sequences, which have one or more reference, to the OMIM database. The number of OMIM entries referenced in SWISSPROT is 6770, as in March 2002 release of the SWISSPROT database. A possible reason for the difference in the numbers of genes could be that more than one genetic disorder entry in OMIM database is associated with a given protein. In the data set used by Muller et al. (*38*) there were 5856 proteins linked to OMIM database entries.

We have used profile-matching method RPS-BLAST that matches a sequence to sequence-profile obtained from structural (PALI – Release 2.2) and functional domain (Pfam – Version 10.0) families. We used stringent e-value cut off of 3×10^{-5} in our search methods to ensure reliability of the domain association. This e-value cut-off has been extrapolated from the one reported by Schaffer et al. (1999) (*29*) as well as based on the benchmarking (N. S. Mhatre, B. Anand and N. Srinivasan, unpublished results) using the database of structure-based sequence alignments of similarly folded proteins. We have used HMMER based procedure against Pfam HMM profiles with an e-value cut off of 10^{-2} to extract reliable domain association. Subsequent to functional/

structural domain identification, all the sequences were subjected to TMHMM2.0 (*49*) in order to assign transmembrane helical regions. The functional and structural domain assignments for human proteome along with other organisms are made publicly available at: http://hodgkin.mbu.iisc.ernet.in/~human. The following sections discuss some of the interesting results derived by analysing this large dataset of human proteins.

OVERALL STRUCTURAL AND FUNCTIONAL DOMAIN ASSIGNMENTS

We could assign a total of 52,297 functional/structural domains to 21,835 (75%) human proteins out of 29,031 proteins encoded in the human genome. We also surveyed for transmembrane regions in human proteins, since this would suggest putative localization of these proteins to the membrane hence, could aid in function prediction. Using TMHMM (*49*) we could identify transmembrane regions in 6777 gene products. Of these, 5424 are found to be present in combination with extracellular or intracellular functional/structural domains. The total number of residues covered in structural/ functional domain and transmembrane region assignments are about 42% of the proteome. These functional/structural domain assignments would indicate probable biochemical functions for the assigned proteins, which could be useful for biological function prediction. A total of 7196 human proteins with no domain assignment, hence with no function or cellular localization information, could form an interesting set for experimental exploration for their properties and biological roles.

The association of gene products with structure can give valuable insights, since structural information provides molecular details of the function of a protein. The structural domain assignment will also help in prioritizing the target for structural genomics consortium by indicating gene products with no structural predictions. With a view to enhance structural information present in human genome, we have used structural information as in PALI profiles that is generated using structure-dependent sequence alignments of a large number of protein domain families, since the incorporation of 3-D structural information could aid in effective detection of remotely related proteins. Using PALI profiles alone, we could associate additional 1191 structural domains to 1076 human proteins that are remotely related.

Furthermore, we tried relating families with unknown structures to known structural families as in SUPFAM database, which was developed by us earlier (*26, 27*) in order to enhance information on the structural content of human proteome. The SUPFAM database relates two or more homologous protein families, of either known or unknown structure, using profiles derived from structure-based sequence alignments. Integrating the relationships derived in SUPFAM we could provide structural information for an additional

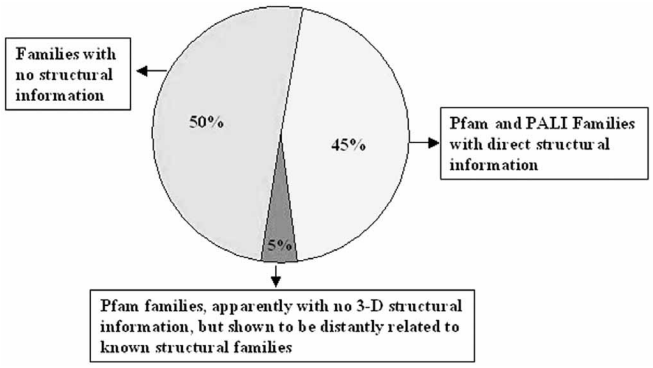

Figure 1. Pie-chart showing the distribution Pfam and PALI families, as identified in the human genome, with respect to their structural information. Some families (about 5%), apparently with unknown structures could be associated with families of known structures using superfamily relationships.

~5% of domain families (Fig. 1). These family assignments would increase known structural content in the genome. A total of 2669 Pfam families are assigned in human genome, of which 1195 Pfam families, have structural information documented in Pfam. Out of 1474 Pfam families, apparently with no structural information, 129 families could be related to a family of known structure in SUPFAM. There are now 1324 families (~50%) with structural information known directly or indirectly through relationships present in SUPFAM (Fig. 1). These 1324 unique families with structural information are present in 40,947 domains, hence would provide further insights into their functions.

DISTRIBUTION OF FUNCTIONAL FAMILIES IN THE HUMAN GENOME

A total number of 52,082 functional domains could be assigned to the human proteins. These assigned functional domains belong to 2669 sequence/functional families of the Pfam database (*42*). We have surveyed for the most commonly occurring Pfam families in human genome. Fig. 2 shows the most frequently occurring functional domain families in the human genome. The most frequently occurring family is the zf-C2H2 family, which is a classical zinc-finger domain with very short length (typically 25 residues). Identification of such a family with short motifs using bioinformatics tools could be unreliable. Hence, we did not consider them in our analysis. The other most frequently occurring globular protein family is protein kinase. It was previously shown that protein kinases occur with typical and atypical combination of domain families in the gene products encoded in human genome. These kinase domain-containing proteins are involved in a wide variety of biological roles (*50*). Among the other most frequently occurring Pfam families, the majority are involved in or in part responsible for protein-protein interactions

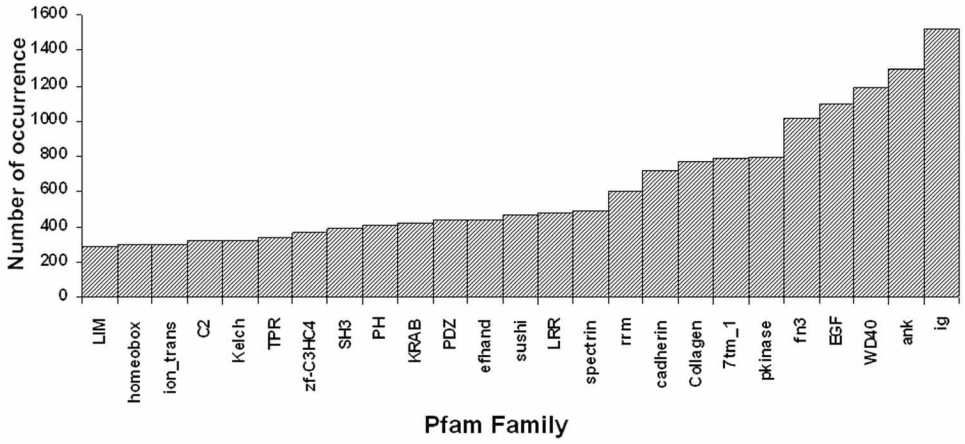

Figure 2. The histogram showing the distribution of most populated protein domain functional families in the human genome. The zf-C2H2 has been excluded from this histogram due to low reliability in assignments of short domains.

(Immunoglobulin, ankyrin repeat, TPR domains), cell attachment adhesion function (Fibronectin, Collagen, Cadherin domains), signalling function (PH, SH3, C2 domains), nucleic acid binding function (zf-C2H2, homoeobox, rrm domains). A considerable number of human proteins are characterized by short lengths, although they match significantly with protein domain families which are typically much longer.

There are 129 functional families, associated in 1144 human proteins, apparently with no structural information but could be associated to distantly related families of known structures using relationship described in SUPFAM. These 1144 proteins have 1184 number of domains. Most of these 129 families correspond to enzymes. From the structural genomics perspective, structure association for 129 Pfam families meant that clues about structure and function could be extended for 1144 proteins.

Atypical Families in the Human Proteome

Some of the Pfam protein families are known to be characteristic of prokaryotic organisms or viruses only. However members of some of these families from human genome could be identified from the current analysis and these families are referred to as atypical families. We could associate 16 bacterial specific families and 18 viral specific families to 41 and 188 human proteins respectively. The list of bacterial and viral specific families, identified in human, along with associated gene products in human genome are listed in Table 1A and 1B respectively The complete list of proteins with the region of Pfam domain assignment is made available at http://hodgkin.mbu.iisc.ernet.in/~human. The assigned domain family includes for example cobalamin biosynthesis protein, minor capsid protein, Bacteriophage lambda head decoration protein. Such functions have not been shown before to be present in humans.

There are two possible explanations that could be drawn in the context of occurrence of the bacterial and viral specific families in human genome. First explanation is that the superfamily relationship exists between the assigned bacterial or viral domain families and the corresponding eurkaryotic domain families as the sequence similarity of these domains with human proteins is low, while significant. These regions in the human proteins could have diverged significantly and sequence data corresponding to these families in other eukaryotes is currently lacking. An alternative possibility is horizontal gene transfer of these bacterial/viral specific families to humans.

Eukaryote Specific Families not Present in the Human Genome

We surveyed the human genome for the occurrence of specific Pfam families, which are known to be present only in eukaryotes. Such eukaryote specific families are known to be involved in specific functions in eukaryotes. Out of 2229 eukaryote specific Pfam families, we could not associate 1054 eukaryotic specific families to the human proteome. Further, we assessed reasons for the absence of these eukaryotic specific families in human genome. Most of these families are organism or lineage specific. Some of them have no known functions and other, as mentioned below, are involved in functions not required or present in humans, hence not identified in the human proteome. The probable reasons for absence of eukaryotic specific Pfam families in human genome are:

(1) Class of toxin families (which also includes snake and scorpion toxins)
(2) Families that are unique to the plant kingdom, like seed storage class of proteins, potato inhibitor and plant disease resistance response protein.
(3) Families which are known to occur only in specific eukaryotes (Yeast & *C. elegans*) like yeast DNA-binding domain, yeast PIR protein repeat, *C. elegans* Sra family integral membrane protein, *C. elegans* integral membrane

Table 1A
List of gene products encoded in the human genome, which are associated with Pfam families with constituent members that are predominantly or exclusively of bacterial origin

Pfam family (Bacterial specific)	ENSEMBL codes for the gene product
ActA Protein	ENSP00000301067
Aspartate-ammonia ligase	ENSP00000263791
Borrelia P83/100 protein	ENSP00000307928
Cobalamin biosynthesis protein CobT	ENSP00000218364
	ENSP00000221166, ENSP00000222271
	ENSP00000236256, ENSP00000252455
	ENSP00000252825, ENSP00000255194
	ENSP00000261722, ENSP00000262518
	ENSP00000265713, ENSP00000273612
	ENSP00000276116, ENSP00000294905
	ENSP00000296126, ENSP00000296755
	ENSP00000299601, ENSP00000302640
	ENSP00000338857, ENSP00000340106
	ENSP00000341529, ENSP00000345250
	ENSP00000345353, ENSP00000345463
DUF164	ENSP00000312056
DUF335	ENSP00000276116
EspF protein	ENSP00000262800
Glutaredoxin 2, C terminal domain	ENSP00000224326
	ENSP00000314404
	ENSP00000345023
Lipopolysaccharide kinase (Kdo)	ENSP00000302185
Borrelia lipoprotein	ENSP00000230165
Mycoplasma MG185/MG260 protein	ENSP00000246043
Neisseria meningitidis TspB protein	ENSP00000309034, ENSP00000329219
Nucleoside H + symporter	ENSP00000281416
Cobalamin biosynthesis protein CobS	ENSP00000251742
DNA-directed RNA polymerase delta subunit	ENSP00000301396
Type specific antigen	ENSP00000319087, ENSP00000342172

protein Srb, *C. elegans* Sre G protein-coupled chemoreceptor and *C. elegans* Srg family integral membrane protein.

This analysis showed that human proteome has eukaryotic specific families (1175) which are involved in eukaryotes-like functions. However, absence of some of the eukaryotic specific families could be explained from the observations that such biochemical functions are undesirable for human or they are highly specific to lower eukaryotes.

Sequence Superfamilies in the Human Proteome

The Pfam families, without known 3-D structure, could be clustered into sequence superfamilies and such superfamily relationships are documented in the SUPFAM database. In the current release of SUPFAM, 96 of the 3904 Pfam families, with no structural information, could be clustered into 39 new potential superfamilies. It is expected that members of all the families in each new potential superfamily would share the same fold and might have gross similarity in their functional properties. These relationships could help in prioritizing the target for structural genomics, since the 3-D structural determination of one of the representative member in each superfamily would result in 39 structures that can serve as framework models. Using these sequence superfamilies information we could identify 18 of the 39 sequence superfamilies in the human genome. The list of these new potential superfamilies with their constituent families identified in human genome is listed in Table 2.

The 18 sequence superfamilies identified in human genome consist of 25 Pfam families, with no known 3-D structure for any of their members. There are 371 domains belonging to these 18 new potential superfamilies that could be assigned to the 367 unique gene products in human genome. Hence, an

Table 1B

List of gene products encoded in the human genome, which are associated with Pfam families with constituent members that are predominantly or exclusively of viral origin

Pfam family (Viral specific)*	ENSEMBL codes for the gene product
Astrovirus capsid protein precursor	ENSP00000216538, ENSP00000342294
Coronavirus non-structural protein NS4	ENSP00000234982, ENSP00000295926
DUF755	ENSP00000225428, ENSP00000318974
	ENSP00000331700
Ebola nucleoprotein	ENSP00000252998
Spumavirus gag protein	ENSP00000229204, ENSP00000262811
	ENSP00000265460, ENSP00000278836
	ENSP00000280333, ENSP00000294256
	ENSP00000296302, ENSP00000299550
	ENSP00000322234, ENSP00000326408
	ENSP00000334414
Geminivirus putative movement protein	ENSP00000319960
Phage tail fibre adhesin Gp38	ENSP00000222330, ENSP00000238823
	ENSP00000246635, ENSP00000247066
	ENSP00000254043, ENSP00000275248
	ENSP00000304994,ENSP00000336604
Herpesvirus BLRF2 protein	ENSP00000233607, ENSP00000238483
Glycoprotein GG/GX	ENSP00000265562, ENSP00000280083
	ENSP00000298229, ENSP00000343897
Herpesvirus polymerase accessory protein	ENSP00000316042
Herpesvirus large structural phosphoprotein UL32	ENSP00000205890, ENSP00000246914
	ENSP00000251041, ENSP00000251819
	ENSP00000259882, ENSP00000262444
	ENSP00000278940, ENSP00000317782
	ENSP00000330326, ENSP00000333262
	ENSP00000337113, ENSP00000339778
	ENSP00000344660
Minor capsid protein VI	ENSP00000344579, ENSP00000268489
Potato leaf roll virus readthrough protein	ENSP00000286760, ENSP00000248610
	ENSP00000251287, ENSP00000329395
	ENSP00000334319, ENSP00000344308
Poxvirus B22R protein	ENSP00000216832, ENSP00000263205
	ENSP00000264160, ENSP00000267260
	ENSP00000276230, ENSP00000279575
	ENSP00000295851, ENSP00000312035
	ENSP00000317898, ENSP00000331396
	ENSP00000342012, ENSP00000344884
TT viral orf 1	ENSP00000221448, ENSP00000235399
	ENSP00000253363, ENSP00000264229
	ENSP00000273628, ENSP00000274514
	ENSP00000295930, ENSP00000301011
	ENSP00000322667, ENSP00000330188
	ENSP00000337194, ENSP00000342705
	ENSP00000343315, ENSP00000344588
	ENSP00000344700, ENSP00000345039
	ENSP00000345947
Tymovirus 45/70Kd protein	ENSP00000345201

*Two Pfam families, viz. Herpes virus major outer envelope glycoprotein (BLLF1) and Totivirus coat protein, are not listed in this table. The numbers of gene products associated to Herpes virus major outer envelope glycoprotein (BLLF1) and Totivirus coat protein Pfam family are 65 and 42 respectively. The list of these gene products is provided on the web site at: http://hodgkin.mbu.iisc.ernet.in/~human).

Table 2
New potential sequence superfamilies that could be associated to the gene product encoded in human proteome

New potential sequence superfamilies	Families in sequence superfamily that occur in human genome	Other families in the sequence superfamily that are not assigned in human genome
1	7 transmembrane receptor (metabotropic glutamate family)	7TM chemoreceptor
2	Amino acid transporter protein	Tryptophan/tyrosine permease family
3	Aromatic-Rich Protein Family	Polyketide cyclase
4	Non-SMC condensin subunit, TBP (TATA-binding protein) – interacting protein 120	None
5	Cobalamin biosynthesis protein	Transferrin binding protein-like solute binding protein
6	DUF608, Amylo-alpha-1,6-glucosidase	Bacterial alpha-L-rhamnosidase, Plant neutral invertase
7	DUF75	DUF774
8	DUF791, Organic Anion Transporter, DUF894, Sugar (and other) transporter	None
9	Glycine cleavage T-protein	Sarcosine oxidase, gamma subunit family
10	HOOK, V-type ATPase	None
11	Lanthionine synthetase C-like protein	NisC-like family
12	MatE	MviN-like protein
13	Methyltransferase, tRNA (Uracil-5-)-methyltransferase	None
14	Nop14-like family	Arsenical pump membrane protein
15	Patched	AcrB/AcrD/AcrF family, Protein export membrane protein (SecD and SecF)
16	Serine carboxypeptidase S28	PS-10 peptidase S37
17	S-antigen protein	Herpesvirus latent membrane protein 1
18	UPF0005	DUF893

experimental structure for 18 domains or proteins one each from these superfamilies could provide templates for interpreting the functions of other members in the superfamily. This results in substantial reduction (from 371 to 18) in the number of 3D-structures to be determined experimentally in order to get clues about their functions experimentally. These superfamilies may be considered as priority targets for structural genomics initiatives in order to improve the coverage of structural information for the human proteins. The nature of the superfamily relationships for some of the new potential sequence superfamilies that are identified in human genome is discussed further.

Patched-like Transport Protein Superfamily

This superfamily consists of three families namely patched domain, ACR_tran and SecD_SecF domain families. The ACR_tran family is an integral membrane protein family whose members are known to be involved in drug resistance in bacteria (*51*). The other family in this superfamily, patched domain, is a receptor for the morphogene sonic hedgehog and transduces hedgehog signals (*52*). This SecD and SecF family consists of various prokaryotic SecD and SecF protein export membrane proteins (*53*). We could identify 16 human proteins with Patched family domain assigned. The functional and structural elucidation of other two families could be could be extended to patched domain because of superfamily relationships.

Transport Superfamily

This superfamily has four families cluster together viz. sugar_tr, OATP_C, DUF791 and DUF894. The sugar_tr family is single-polypeptide capable only of transporting small solutes, such as sugar, in response to chemiosmotic ion gradients and lies in uniporter-symporter-antiporter family (*54*). OATP_C is eukaryotic Organic-Anion-Transporting Polypeptides that vary in tissue distribution and substrate specificity (*55*). The functions of DUFs (Domains of Unknown Function) are not known. We could associate sugar_tr and OATP_C domains to 85 and 16 gene products respectively.

Methyltransferase Superfamily

This superfamily constitutes Methyltransf_4 and tRNA_U5-meth_tr Pfam families. Both families have methyltransferase activity, however, the tRNA_U5-meth_tr family is

involved in methylation of t-RNAs (*56*). We could identify 1 and 5 human homologues of Methyltransf_4 and tRNA_U5-meth_tr respectively.

Glucosidase superfamily

The GDE_C and DUF608 Pfam families could be clustered in this superfamily. The GDE_C family is glycogen branching enzyme and has glucosidase activity (*57*). We could identify GDE_C and DUF608 in 3 and 2 human proteins respectively. From, this relationship it could be suggested that DUF608 might have glucosidase-like activity.

DISTRIBUTION OF PROTEIN DOMAIN FOLDS AND STRUCTURAL SUPERFAMILIES IN THE HUMAN GENOME

The 3-D information provides precise molecular details about the function of the protein. The association of gene products encoded in human genome to 3-D structures would assist in providing further insights into their function. The databases of protein structures in which domains with similar 3-D architecture are grouped together could be used for such structural analysis. We have used PALI database derived from SCOP for the present analysis. SCOP classifies protein domain having high sequence and structural similarity into families. The families are grouped in superfamilies when they share similar functional features and have an evolutionary common ancestor. Superfamilies are grouped in fold when major secondary structures are topologically equivalent with similar topological connectivity. The assignment of structural domains to the proteins would aid in the investigation of the preponderance of superfamilies and fold in the human genome.

Using the various search procedures we could associate 38,017 structural domains to 16,459 human proteins either directly or by using the sequence superfamily relationships as described in SUPFAM. Further, we classified these domain assignments at the level of fold or superfamilies to understand the most commonly used function present in human genome.

We analyzed the most commonly occurring superfamilies in human proteome. The Figure 3 shows the top few superfamilies along with their extent of representation in the human genome. This distribution of superfamilies is similar to the one obtained by Muller et al. (*38*). The most commonly occurring superfamily is C2H2 zinc finger, followed by immunoglobulin domain. Because of the short length of C2H2 zinc-finger domain and associated low complexity region, there is bias in identification of these domains. Hence, all the gene products having this domain might not have zinc-finger like function and we excluded them from our present analysis. P-loop containing nucleotide triphosphate hydrolases domain is the next most represented superfamily and it is involved in many different critical biological functions such as cell growth, differentiation, repair and modification of DNA, transcription, etc. This superfamily comprises various ATPases and GTPases that are essential for cell survival. For example GTPases include elongation factors, Gα subunit of the heterotrimeric G-proteins that are absolutely critical in major cellular processes. The other superfamilies among frequently occurring superfamily are involved in various functions in the cell as cellular signalling (Protein kinase-like, PH domain-like), cell adhesion (Cadherin, Fibronectin type III), nucleic acid binding function (RNA-binding domain). Interestingly, 'Family A-G protein-coupled receptor-like' superfamily that consists of many receptors as other most populous superfamilies. The complete list of structural superfamilies that occur in human genome with their respective frequency of occurrence in human genome is provided at http://hodgkin.mbu.iisc.ernet.in/~human.

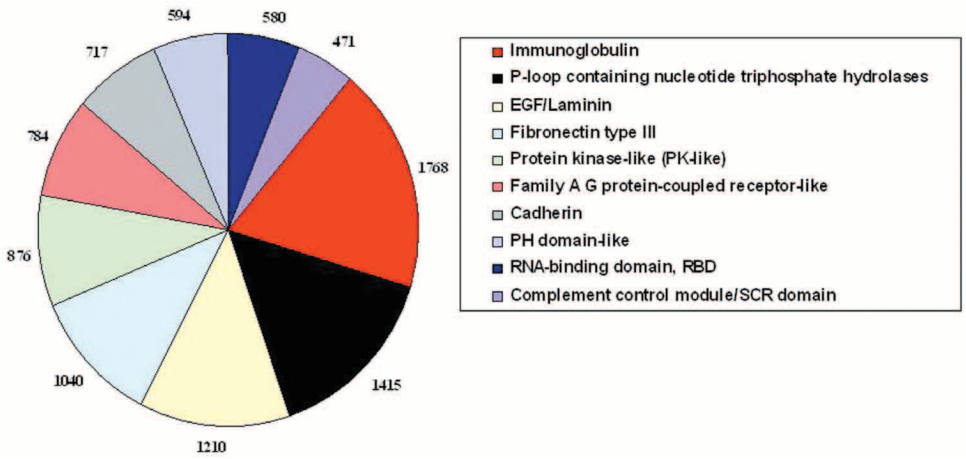

Figure 3. Population distribution of the most populated superfamilies of known 3-D structure in the human genome. The C2H2 zinc finger has been excluded from this distribution.

Figure 4 shows population distribution of few most populated folds, which occur in the human proteome. Figure 5 shows the 3-D folding patterns in the most populated folds. The C2H2 and C2HC zinc finger is the most frequent occurring fold in human genome. For the reasons mentioned before, we have excluded C2H2 and C2HC zinc finger from this analysis. Ferredoxin-like fold has the highest number of superfamilies in the human proteome as well as in SCOP. However, 16 of the superfamilies occur in the human proteome out of the currently known 36 superfamilies in the ferredoxin fold. Ribonuclease H-like motif fold has six out of currently known seven superfamilies in the human proteome. Except the superfamily of hypothetical protein MTH1175 from methanobacterium, all other superfamilies of Ribonuclease H-like motif occur in the human proteome. These superfamilies are Actin-like ATPase domain, Creatinase/prolidase N-terminal domain, Ribonuclease H-like, translational machinery components, DNA repair protein Muts domain II and Methylated DNA-protein cysteine methyltransferase domain. This could be expected as the nucleic acid binding/related superfamilies are highly represented in the human proteome. The complete list of protein structure folds that occur in human genome with their respective frequency of occurrence in human genome is provided at http://hodgkin.mbu.iisc.ernet.in/~human.

ASSIGNMENT OF DOMAIN FAMILIES TO THE PROTEINS IN OMIM DATABASE OF HUMAN DISEASES: NEW DOMAIN ASSIGNMENTS AND THEIR IMPLICATIONS

The sequence to profile matching procedure described in the Methods section resulted in the association of at least one functional domain family in Pfam database to the 4864 proteins of SWISSPROT database linked to OMIM entries (77.8% of the total of 6257 proteins in the OMIM database). The remaining 1393 disease-related proteins could not be associated to any functional or structural domain family. Hence these proteins could be high priority targets in structural genomics to provide further insights into the molecular basis of the function of these proteins.

These 4864 proteins contain 8431 functional and structural domains from 1288 Pfam families. It is important to note that 6491 domains out of 8431 domains could be linked to 802 Pfam families with known structural information. In terms of the amino acids coverage in these domains about 51% of the amino acids in the proteins are in the functional or structural domain (*58*) assigned regions in these 4864 proteins.

Figure 6 shows the distribution of the domains in the top 15 most populous families in the proteins, these families contain 2551 domains which is about 30% of all assigned domains. Protein kinase is the most frequently occurring domain family in the human disease proteins. Among the top 15 most populous families 14 have known structural information. The most populous structural superfamily that is assigned to the proteins is P-loop containing nucleotide triphosphate hydrolases and this has 361 domains in it. The largest representations in the P-loop superfamily come from the domain families like Ras, helicase_c, and DEAD. The list of highly populated superfamilies has much in common with the analogous list generated by Muller et al. (*38*). Much of these highly represented superfamilies are associated with regulatory roles in development, differentiation and proliferation.

Further analysis revealed that there are 33 proteins that have been assigned additional functional domains apart from previously assigned functional domains. These 33 proteins are listed in Table 3. These newly assigned domains may play a significant role in furthering our understanding of overall functions of these proteins.

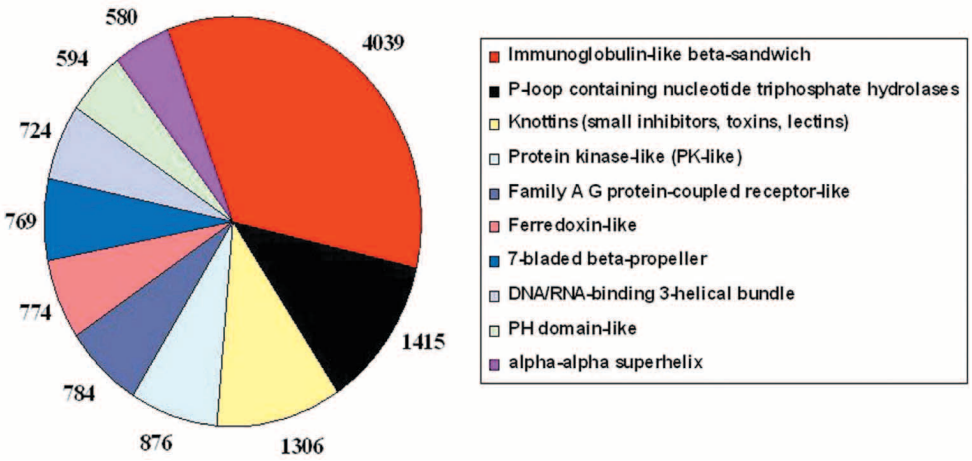

Figure 4. The extent of occurrence of the most commonly occurring protein structure folds associate to gene product encoded in the human genome. The C2H2 zinc finger fold has been excluded from this distribution.

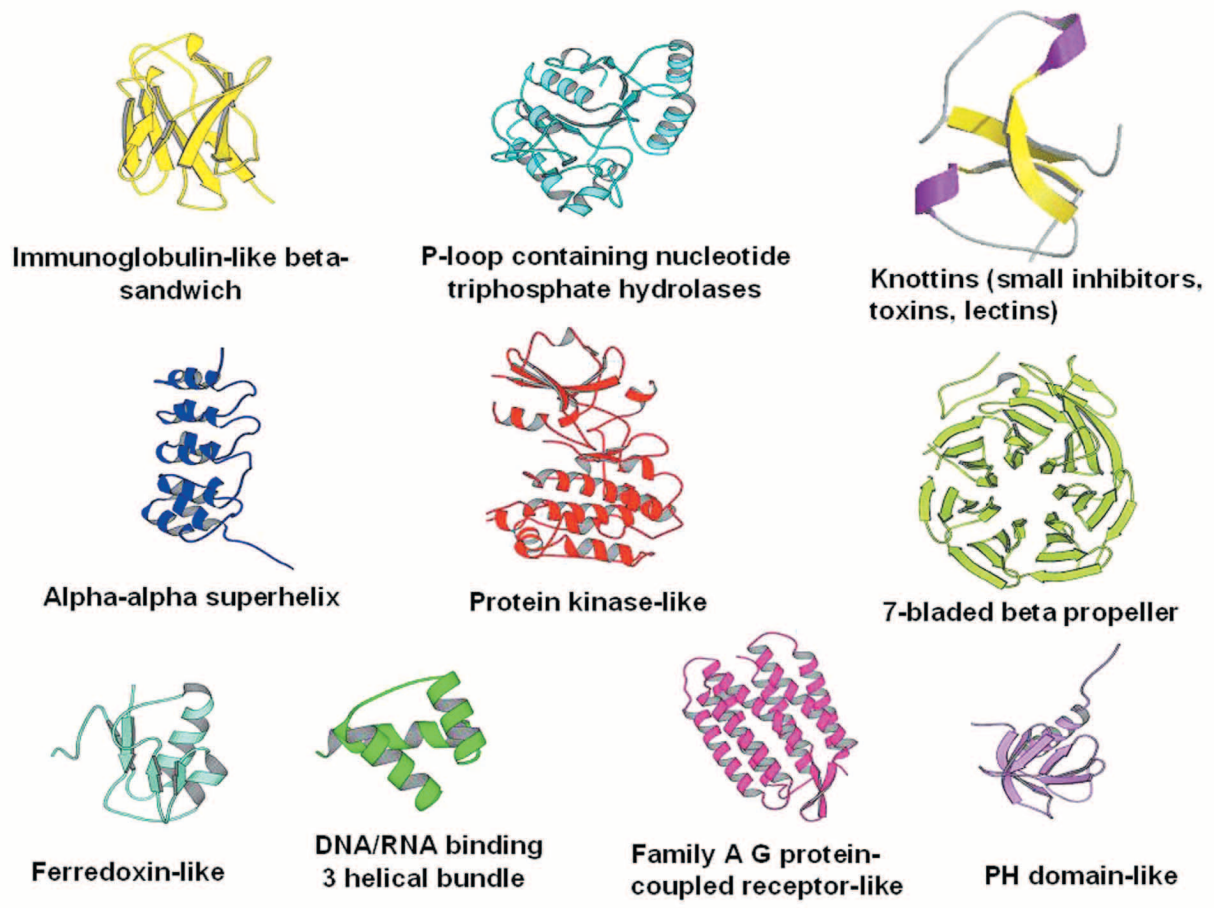

Figure 5. Cartoon representation of 3-D folds of protein domains that are most populated in the human genome. The population of each of these folds is indicated in Fig. 4. The protein structure representations in this figure are generated using the Setor software (*59*).

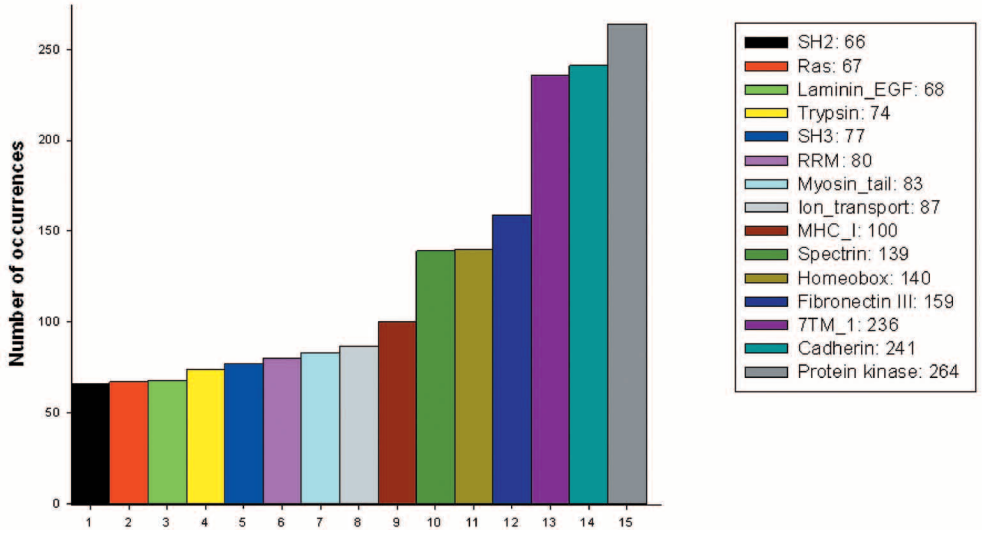

Figure 6. Population distribution of most populated protein domain families present in the human proteins that are potentially linked to genetic disorders and diseases.

Table 3

List of proteins that are potentially linked to genetically inherited human diseases, with the newly assigned functional domains from the present analysis

Protein	Description adopted from OMIM database/Swiss	Previously assigned Pfam domains	Newly assigned Pfam domains	Homologue(s) in other model organisms
Acrosin	Acrosin deficiency leads to infertility in males	Trypsin	Minor capsid protein family VI	Mouse
Myelin transcription factor 1	Proteolipid binding protein (PLPB1)	Zinc finger C2HC type	Heat shock protein 90 (HSP90)	Mouse and Rat
Glutamate receptor (GRIN2C)	Mutation in the gene that codes for the protein lead to defective motor coordination in mice	Ligand gated ion channel	Bacterial extracellular solute binding protein (family 3)	Mouse
Annexin	Implicated in autoimmune diseases	Annexin	dwarfin	Mouse and fruit fly
Son of sevenless protein homolog 1, SOS1	Promotes the exchange of Ras-bound GDP by GTP	PH, RasGef, Rhogef & Rasgefn	dwarfin	Mouse
Drebrin, Developmentally regulated brain protein	Drebrins might play some role in cell migration, extension of neuronal processes and plasticity of dendrites, respectively	Cofilin/tropomyosin-type actin-binding protein	Ezrin/Radixin/Moesin (ERM) family	Mouse
Myosin heavy chain 11, MYH11	Implicated in myeloid leukemia	Myosin_head, myosin_tail & myosin_N	ERM family	Mouse
SWI/SNF related, actin dependent regulator of chromatin	Mammalian SNF/SWI complexes are ATP dependent chromatin remodelling enzymes that are implicated in gene expression, cell cycle and oncogenesis	Snf2 family N-terminal domain, Helicase conserved C-terminal domain & bromodomain	dwarfin	Mouse
Small nuclear ribonucleoprotein, SNRP70	Implicated in auto-immune diseases	RRM (RNA recognition motif)	TT viral orf 1	Fruit fly, Yeast and Mouse
Glutamate receptor, Ionotropic, GRIK4	Acts as neurotransmitter in the mammalian nervous system	Ligand gated ion channel & anf_receptor (Receptor family ligand binding region)	Bacterial solute binding proteins (family 3)	Rat
Histone deactylase 5, HDAC5	It is implicated in colon cancer	Histone deactylase	ERM family	Mouse
Nucleolin, NCL	It is acidic phosphoprotein of exponentially growing cells	RRM	Astro virus capsid family	Mouse
Keratin, type II cytoskeletal 6A	Implicated in Jadassohn-Lewandowsky syndrome	Intermediate filament protein	Keratin	Bovine

(*continued overleaf*)

Table 3
(*continued*)

Protein	Description adopted from OMIM database/Swiss	Previously assigned Pfam domains	Newly assigned Pfam domains	Homologue(s) in other model organisms
Brain-specific angiogenesis inhibitor 1	It is suspected to be linked to progression of glioma to glioblastoma	7TM_2 (transmembrane helix family 2), Latrophilin/ CL-1-like GPS domain, Hormone receptor domain, Thrombospondin type 1 domain	CAP	Mouse and Rat
Upstream binding transcription factor UBTF	Required for expression of 18S, 28S and 58S ribosomal RNAs	HMG (high mobility group) box	Astrovirus capsid protein	Mouse, Rat and Frog
Adisintegrin and metalloproteinase 29	Involved in cell-cell, cell-matrix interactions linked to fertilisation, muscle development and neurogenesis	Reprolysin family propeptide, disintegrin & reprolysin	TYA transposon protein	Mouse
Phosphatidylinositol glycan (PIGL)	GPI12 is ortholog of human PIGL in yeast. Disruption of GPI12 in yeast resulted in lethal phenotype		DUF158	Yeast and Rat
Tumor susceptibility gene 101, TSG101	TSG101 is mutated in high frequency in breast cancer and there was further finding that defects in TSG101 occurs during breast cancer tumorogenesis		DUF164	Mouse
VSKI avian sarcoma viral oncogene homolog; SKI	Implicated in oncogenesis	SKI/SNO/DAC family	ERM family	Mouse and Frog
3-hydroxyl-3-methyl glutaryl CoA reductase (HMGCAR)	Regulated expression of HMG_CoA reductase has a critical development in providing spatial information to guide migrating primordial germ cells	Hydroxymethylglutaryl-coenzyme A reductase	patched	Fruit fly, Mouse, Rat and Frog
FAT tumor suppressor	Important in mammalian development process and cell proliferation	Cadherin, EGF-like domain and laminin domain	dwarfin	Mouse and Rat
Myeloid/Lymphoid mixed lineage leukemia	Implicated in myeloid leukemia	PHD-finger, SET & Zinc finger-CXXC	CAP	Mouse
Utrophin, UTRN	Implicated in muscular dystrophy	Calponin homology (CH) domain, Spectrin & Zinc finger-ZZ	spectrin	Mouse

(*continued overleaf*)

Table 3
(*continued*)

Protein	Description adopted from OMIM database/Swiss	Previously assigned Pfam domains	Newly assigned Pfam domains	Homologue(s) in other model organisms
Tumor necrosis factor ligand superfamily, TNFSF6	Implicated in tumor	TNF(Tumor Necrosis Factor) family	CAP	Mouse and Rat
POU-domain, class 4, POU4F1	It is class 4 POU domain containing transcription factor highly expressed in the developing sensory nervous system and in cells of B & T-lymphocytic lineages	Homeobox & Pou domain	Prion	Mouse, Rat and Chick
Glutamate receptor, GRIA4	The postsynaptic actions of glu are mediated by a variety of receptors that are named according to their selective agonists.	Anf_receptor & ligand gated ionchannel	Bacterial solute binding protein (family 3)	Mouse and Rat
Splicing factor, P and Q rich; SFPQ	Essential pre-mrna splicing factor required early in spliceosome formation.	RRM	Dwarfin and apolipoprotein	Fruit fly
Ryanodine receptor 1	It is linked to central core disease of muscle	MIR, RIH domain, RYR, SPRY & Ion transport protein	DUF236 and Nucleosome assembly protein	Rabbit and Pig
Zonadhesion, ZAN	ZAN is a sperm membrane protein that binds zona pellucida of the egg in a species-specific manner.	MAM, Trypsin Inhibitor like cysteine rich domain (TIL), TILa domain & von Willebrand factor type D domain	Syndecan	Mouse, Pig and Rabbit
Ankyrin 2	Attaches integral membrane proteins to cytoskeletal elements. Also bind to cytoskeletal proteins.	Zu5, Ank repeat & DEATH	ion-channel transmembrane region	Rat, Mouse, Fruit fly and Worm
WAS2_HUMAN	Wiskott-Aldrich syndrome protein family member 2. Downstream effector molecules involved in the transmission of signals from tyrosine kinase receptors and small GTPases to the actin cytoskeleton.	WH2__	Herpes_gg	–
T101_HUMAN	Tumor susceptibility gene 101 protein		duf 164	Mouse
UDP-glucose 4-epimerase	Galactose epimerase deficiency, GALR deficiency	Epimerase	3-beta hydroxysteroid dehydrogenase/isomerase family	Fruit fly and Rat

OUTLOOK

Using various methods of domain association we could associate at least one domain to about 75% of gene products in the human genome. Interestingly, the assignments of remote homologues to the human proteins revealed the occurrence of some of the viral and bacterial specific proteins in the human genome. Among most commonly occurring functional family, Protein kinases is one of the most frequently occurring domains, and the P-loop containing nucleotide triphosphate hydrolases is the one of the most represented superfamily. The assignment of 1184 domains to families with apparently no structural information to structural families would aid in the prioritization of targets for structural genomics of human genome. The assignment of new domains in addition to previously identified domains to the proteins possibly linked to genetically inherited human diseases could form a basis for the experimental verification of the roles of these domains as well as the molecular basis of disease.

ACKNOWLEDGEMENTS

This research is supported by the award of Senior Fellowship to N.S. by the Wellcome Trust, London as well as by the computational genomics initiative supported by the Department of Biotechnology, New Delhi. S.B. and S.B.P. are supported by the Wellcome Trust, London and CSIR, New Delhi respectively.

REFERENCES

1. Lander, E. S. et al. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.
2. Ventor, J. C. et al. (2001) The sequence of the human genome. *Science* **291**, 1304–1351.
3. Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M. and Yuan, Y. (1998) Predicting function: from genes to genomes and back. *J. Mol. Biol.* **283**, 707–725.
4. Teichmann, S. A., Park, J. and Chothia, C. (1998) Structural assignments to the Mycoplasma genitalium proteins show extesive gene duplication and domain rearrangements. *Proc. Natl. Acad. Sci. USA* **95**, 14658–14663.
5. Pearl, F. M., Lee, D., Bray, J. E., Buchan, D. W., Shepherd, A. J. and Orengo, C. A. (2002) The CATH extended protein-family database providing structural annotations for genome sequences. *Protein Sci.* **11**, 233–244.
6. Aravind, L., Mazumder, R., Vasudevan, S. and Koonin, E. V. (2002) Trends in protein evolution inferred from sequence and structure analysis. *Curr. Opin. Struct. Biol.* **12**, 392–399.
7. Harrison, P. M. and Gerstein, M. (2002) Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. *J. Mol. Biol.* **318**, 1155–1174.
8. Huynen, M., Snel, B., Lathe, W. 3rd and Bork, P. (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.* **10**, 1204–1210.
9. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
10. Pearson, W. R. and Lipman, D. J. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85**, 2444–2448.
11. Murzin, A. G. and Bateman, A. (1997) Distant homology recognition using structural classification of proteins. *Proteins* **Suppl 1,** 105–112.
12. Chothia, C. and Lesk, A. M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826.
13. Chothia, C. and Gerstein, M. (1997) Protein evolution. How far can sequences diverge? *Nature* **385**, 579–581.
14. Gerstein, M. (1998) How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Fold. Des.* **3**, 497–512.
15. Huynen, M., Doerks, T., Eisenhaber, F., Orengo, C., Sunyaev, S., Yuan, Y. and Bork, P. (1998) Homology-based fold predictions for *Mycoplasma genitalium* proteins. *J. Mol. Biol.* **280**, 323–326.
16. Hegyi, H. and Gerstein, M. (1999) The relationship between protein structure and function a comprehensive survey with application to the yeast genome. *J. Mol. Biol.* **288**, 147–164.
17. Kelley, L. A., MacCallum, R. M. and Sternberg, M. J. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.* **299**, 499–520.
18. Fischer, D. and Eisenberg, D. (1999) Predicting structures for genome proteins. *Curr. Opin. Struct. Biol.* **9**, 208–211.
19. Orengo, C. A., Todd, A. E. and Thornton, J. M. (1999) From protein structure to function. *Curr. Opin. Struct. Biol.* **9**, 374–382.
20. Fetrow, J. S., Siew, N., Di Gennaro, J. A., Martinez-Yamout, M., Dyson, J. H. and Skolnick, J. (2001) Genomic-scale comparison of sequence-and structure-based methods of function prediction: Does structure provide additional insight? *Protein Sci.* **10**, 1005–1014.
21. Gribskov, M., McLachlan, A. D. and Eisenberg, D. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA* **84**, 4355–4358.
22. Krogh, A., Brown, M., Mian, I. S., Sjolander, K. and Haussler, D. (1994) Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.* **235**, 1510–1531.
23. Rychlewski, L., Zhang, B. and Godzik, A. (1998) Fold and function predictions for *Mycoplasma genitalium* proteins. *Fold Des.* **3**, 229–238.
24. Bork, P. and Gibson, T. J. (1996) Applying motif and profile searches. *Methods Enzymol.* **266**, 162–184.
25. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhanng, Z., Miller, W. and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
26. Pandit, S. B., Gosar, D., Abhiman, S., Sujatha, S., Dixit, S. S., Mhatre, N. S., Sowdhamini, R. and Srinivasan, N. (2002) SUPFAM-a database of potential protein superfamily relationships derived by comparing sequence-based and structure-based families: implications for structural genomics and function annotation in genomes. *Nucl Acids Res.* **30**, 289–293.
27. Pandit, S. B., Bhadra, R., Gowri, V. S., Balaji, S., Anand, B. and Srinivasan, N. (2004) SUPFAM: A database of sequence superfamilies of protein domains. *BMC Bioinformatics* **5**, 28.
28. Namboori, S., Mhatre, N., Sujatha, S., Srinivasan, N. and Pandit, S. B. (2004) Enhanced functional and structural domain assignments using remote similarity detection procedures for proteins encoded in the genome of *Mycobacterium tuberculosis* H37Rv *J. of Biosci.* (in press).
29. Schaffer, A. A., Wolf, Y. I., Ponting, C. P., Koonin, E. V., Aravind, L. and Altschul, S. F. (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics* **15**, 1000–1011.

30. Eddy, S. R. (1998) Profile hidden Markov models. *Bioinformatics* **14**, 755–763.
31. Muller, A., MacCallum, R. M. and Sternberg, M. J. (1999) Benchmarking PSI-BLAST in genome annotation. *J. Mol. Biol.* **293**, 1257–1271.
32. Lindahl, E. and Eloffson, A. (2000) Identification of related proteins on family, superfamily and fold level. *J. Mol. Biol.* **295**, 613–625.
33. Jones, D. T. (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* **287**, 797–815.
34. Bowie, J. U., Luthy, R. and Eisenberg, D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**, 164–170.
35. Zdobnov et al. (2002) Comparative genome and proteome analysis of Anopheles gambiae and Drosophila melanogaster. *Science* **298**, 149–159.
36. Snel, B., Bork, P. and Huynen, M. A. (2002) The identification of functional modules from the genomic association of genes. *Proc. Natl. Acad. Sci. USA* **99**, 5890–5895.
37. Lin, J., Qian, J., Greenbaum, D., Bertone, P., Das, R., Echols, N., Senes, A., Stenger, B. and Gerstein, M. (2002) GeneCensus: genome comparisons in terms of metabolic pathway activity and protein family sharing. *Nucl. Acids Res.* **30**, 4574–4582.
38. Muller, A., MacCallum, R. M. and Sternberg, M. J. (2002) Structural characterization of the human proteome. *Genome Res.* **12**, 1625–1641.
39. Koonin, E. V., Wolf, Y. I. and Arvind, L. (2000) Protein fold recognition using sequence profiles and its application in structural genomics. *Adv. Prot. Chem.* **54**, 245–275.
40. Frishman, D., Albermann, K., Hani, J., Heumann, K., Metanomski, A., Zolner, A. and Mews, H. W. (2001) Functional and structural genomics using PEDANT. *Bioinformatics* **17**, 44–57.
41. Iliopoulos, I., Tsoka, S., Andrade, M. A., Jansen, P., Audit, B., Tramontano, A., Valencia, A., Leroy, C., Sander, C. and Ouzounis, C. A. (2001) Genome sequences and great expectations. *Genome Biol.* **2**, 0001.1–0001.3.
42. Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L. and Sonnhammer, E. L. L. (2000) The Pfam protein families database. *Nucl. Acids Res.* **28**, 263–266.
43. Balaji, S., Sujatha, S., Kumar, S. S. C. and Srinivasan, N. (2001) PALI-a database of Phylogeny and ALIgnment of homologous protein structures. *Nucleic Acids Res.* **29**, 61–65.
44. Gowri, V. S., Pandit, S. B., Karthik, P. S, Srinivasan, N. and Balaji, S. (2003) Integration of related sequences with protein three-dimensional structural families in an updated version of PALI database. *Nucl. Acids Res.* **31**, 486–488.
45. Murzin, A. G., Brenner, S. E., Hubbard, T. and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.
46. Hubbard et al. (2002) The Ensembl genome database project. *Nucl. Acids Res.* **30**, 38–41.
47. Antonarakis, S. E. and McKusick, V. A. (2000) OMIM passes the 1000-disease-gene mark. *Nat. Genet.* **25**, 11.
48. Bairoch, A. and Apweiler, R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucl. Acids Res.* **28**, 45–48.
49. Sonnhammer, E. L. L., Von Heijne, G. and Krogh, A. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. In *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology, Menlo Park, California.* (Glasgow, J., Littlejohn, T., Major, F., Lathrop, R., Sankoff, D., Sensen, C. eds.), pp. 175–182.
50. Krupa, A. and Srinivasan, N. (2002) The repertoire of protein kinases encoded in the draft version of the human genome: atypical variations and uncommon domain combinations. *Genome Biol.* **3**, 0066.1–0066.14.
51. Okusu, H., Ma, D. and Nikaido, H. (1996) AcrAB efflux pump plays a major role in the antibiotic resistance phenotype of Escherichia coli multiple-antibiotic-resistance (Mar) mutants. *J. Bacteriol.* **178**, 306–308.
52. Hooper, J. E. and Scott, M. P. (1989) The Drosophila patched gene encodes a putative membrane protein required for segmental patterning. *Cell* **59**, 751–765.
53. Arkowitz, R. A. and Wickner, W. (1994) SecD and SecF are required for the proton electrochemical gradient stimulation of preprotein translocation. *EMBO J.* **13**, 954–963.
54. Pao, S. S., Paulsen, I. T. and Saier, M. H. (1998) Major facilitator superfamily. *Microbiol. Mol. Biol. Rev.* **62**, 1–34.
55. Tamai, I., Nezu, J., Uchino, H., Sai, Y., Oku, A., Shimane, M. and Tsuji, A. (2000) Molecular identification and characterization of novel members of the human organic anion transporter (OATP) family. *Biochem. Biophys. Res. Commun.* **273**, 251–260.
56. Johansson, M. J. and Bystrom, A. S. (2002) Dual function of the tRNA (m(5)U54) methyltransferase in tRNA maturation. *RNA* **8**, 324–335.
57. Nakayama, A., Yamamoto, K. and Tabata, S. (2001) Identification of the catalytic residues of bifunctional glycogen debranching enzyme. *J. Biol. Chem.* **276**, 28824–28828.
58. Primo-Parmo, S. L., Sorenson, R. C., Teiber, J. and La Du, B. N. (1996) The human serum paraoxonase/arylesterase gene (PON1) is one member of a multigene family. *Genomics* **33**, 498–507.
59. Evans, S. V. (1993) SETOR: hardware lighted three-dimensional solid model representations of macromolecules. *J Mol. Graph.* **11**, 134–138.