EVOLUTIONARY CLASSIFICATION OF PROTEIN DOMAINS: FROM REMOTE HOMOLOGY TO FAMILY

APPROVED BY SUPERVISORY COMMITTEE

NOTE: The top line is for the Supervising Professor's name. There should be as many lines as there are members of the committee. All signatures must be original and in ink. Adjust "Approved by Supervisory Committee" line upward if the committee list is very large.

First Name Last Name, credentials

Nick Grishin

Jose Rizo-Rey

Luke Rice

Diana Tomchick

DEDICATION

I would like to thank the committee members and my mentor for their support and suggestions in my graduate study. Especially, I am grateful to Nick, who gave me such a great opportunity to participate in the project and advise all students properly according to their styles. I feel lucky enough to have spent my last 7 years here and had freedom to explore the directions I am interested in.

I will not be here today without the support from my family and my girlfriend. Their understanding and patience are priceless. Thanks for enduring my selfishness from time to time.

EVOLUTIONARY CLASSIFICATION OF PROTEIN DOMAINS: FROM REMOTE HOMOLOGY TO FAMILY

by

YUXING LIAO

DISSERTATION

Presented to the Faculty of the Graduate School of Biomedical Sciences

The University of Texas Southwestern Medical Center at Dallas

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

The University of Texas Southwestern Medical Center at Dallas

Dallas, Texas

Degree Conferral December, 2017

EVOLUTIONARY CLASSIFICATION OF PROTEIN DOMAINS: FROM REMOTE HOMOLOGY TO FAMILY

Publication No. ________________

Yuxing Liao, Doctor of Philosophy

The University of Texas Southwestern Medical Center at Dallas, 2017

Supervising Professor: Nick V. Grishin, Ph.D.

Understanding the evolution of a protein, including both close and distant relationships, often reveals insight into its structure and function. A protein domain classification splits protein into domains and organizes them according to their evolutionary history. Existing classification databases fall back the speed of protein structure determination and do not include some known homologous relationships. I have participated in creating a hierarchical evolutionary classification of all proteins with experimentally determined spatial structures and developed a website for easy access and searches with keyword, sequence or structure (http://prodata.swmed.edu/ecod). ECOD (Evolutionary Classification Of protein Domains) is distinct from other structural classifications in that it groups domains primarily by evolutionary relationships (homology), rather than topology (or fold). Our database uniquely emphasizes distantly related homologs that are difficult to detect, and thus catalogs the largest number of evolutionary relationships among structural domain classifications. Placing distant homologs together underscores the ancestral similarities of these proteins and draws attention to the most important regions of sequence and structure, as well as conserved functional sites. The classification is assisted by an automated pipeline that classifies the most of new structures in Protein Data Bank weekly. This synchronization uniquely distinguishes ECOD among all protein classifications. For proteins that lack confident results from the automatic pipeline, I rely on information from literature, sequence and structure similarity scores, visual comparison and experience to classify them manually. I document the manual curation process in detail with an example of the remote homology between an autoproteolytic domain found in GPCR-Autoproteolysis Inducing domain, ZU5 and nucleoporin98. ECOD also recognizes closer relationships at the family level, initially with Pfam families. However, existing family databases do not cover all structures and disagree with ECOD in terms of domain definition and boundary. I generate multiple sequence alignment and profile for domains in the same family with structural information and demonstrate that the alignment quality is similar to manually checked Pfam seed alignments. I compare ECOD family profiles with Pfam and Conserved Domain Database and discuss about the improvement of domain boundary over known families and the dominance of small families in new families.

TABLE OF CONTENTS

PRIOR PUBLICATIONS

1. Kinch L, Yong Shi S, Cong Q, Cheng H, Liao Y, Grishin NV. CASP9 assessment of free modeling target predictions. Proteins. 2011;79 Suppl 10:59-73.
2. Chen B, Brinkmann K, Chen Z, Pak CW, Liao Y, Shi S, et al. The WAVE regulatory complex links diverse receptors to the actin cytoskeleton. Cell. 2014;156(1-2):195-207.
3. Cheng H*, Schaeffer RD*, Liao Y*, Kinch LN, Pei J, Shi S, et al. ECOD: an evolutionary classification of protein domains. PLoS Comput Biol. 2014;10(12):e1003926.
4. Liao Y, Pei J, Cheng H, Grishin NV. An ancient autoproteolytic domain found in GAIN, ZU5 and Nucleoporin98. Journal of molecular biology. 2014;426(24):3935-45.
5. Cheng H*, Liao Y*, Schaeffer RD*, Grishin NV. Manual classification strategies in the ECOD database. Proteins. 2015;83(7):1238-51.
6. Ovchinnikov S, Kinch L, Park H, Liao Y, Pei J, Kim DE, et al. Large-scale determination of previously unsolved protein structures using evolutionary information. Elife. 2015;4:e09248.
7. Schaeffer RD, Kinch LN, Liao Y, Grishin NV. Classification of proteins with shared motifs and internal repeats in the ECOD database. Protein science : a publication of the Protein Society. 2016;25(7):1188-203.
8. Schaeffer RD, Liao Y, Cheng H, Grishin NV. ECOD: new developments in the evolutionary classification of domains. Nucleic acids research. 2017;45(D1):D296-D302.

* These authors contributed equally to this work.

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

LIST OF DEFINITIONS

HMM: Hidden Markov Model

CDD: Conserved Domain Database

NCBI: National Center for Biotechnology Information

SCOP: Structural Classification of Proteins

CATH: Class, Architecture, Topology, Homology

PDB: Protein Data Bank

PSSM: position specific scoring matrix

SVM: Support Vector Machine

SQL: Structured Query Language

MRP: Mitochondrial RNA-binding Protein

GPCR: G-protein-coupled receptors

GPS: GPCR proteolysis site

GAIN domain: GPCR autoproteolysis inducing domain

PKD: polycystic kidney disease

Nup98: Nucleoporin98

CILP: cartilage intermediate layer protein

DD: death domain

ZO-1: zonula occluden-1

BAI3: brain angiogenesis inhibitor 3.

ARM repeat: Armadillo repeat

HAD: haloacid dehydrogenase

SAM: Sterile alpha motif

H2TH domain: helix-2turn-helix domain

LGA: Local-Global Alignment

CHAPTER ONE

Introduction

With recent technology advances in the field of genome sequencing and protein structure determination, the numbers of protein sequence and structure have been increasing at an exponential rate (Rose et al., 2015; Stephens et al., 2015). In the face of a deluge of biological data, a classification system helps to reduce the complexity and redundancy and to provide insights of relationship between proteins. As proteins are the product of evolution, classification is most meaningful if it groups related proteins together under the guidance of evolution.

The evolutionary relationship between proteins can be studied by comparing amino acid sequences, starting from sequence identity. Close homologs *i.e.* proteins that did not diverge too long ago in evolution, have less time to accumulate changes in sequence and thus share high sequence similarity. By modeling the amino acid substitution rates empirically, methods like BLAST can readily search and score proteins with high sequence identities to the query (Altschul et al., 1997). As the evolutionary distance between proteins increases, more sensitive sequence homology detection method is needed. It has been demonstrated that using the position specific conservation information from a group of related proteins, or a profile, substantially improves the performance of homology detection, since it distinguishes positions that are more important for maintaining the function or structure of the protein. There are different ways to model the profile (Altschul et al., 1997; Eddy, 1998; Marchler-Bauer et al., 2011) and methods for aligning sequence to profile and profile to profile

(Altschul et al., 1997; Eddy, 2011; Soding, 2005). However, sequence information alone may be insufficient for remote homologs. Various evolutionary events such as duplication, deterioration, fusion, and mutation can alter both the structure and function of homologous proteins; sometimes to the extent that their ancestry can be difficult to discern by sequence similarity or they may adopt fold changes (Grishin, 2001a, 2001b; A. N. Lupas, Ponting, & Russell, 2001). Although structure diverges more slowly than sequence, structural similarity is potentially confounded by the possibility of analogy, or convergent evolution (Cheng, Kim, & Grishin, 2008; Krishna & Grishin, 2004). Therefore, judicious consideration of all aspects of information is necessary, sometimes by an expert.

Many protein classifications are currently available. Comprehensive sequence-based classifications such as Pfam (Finn et al., 2014) and CDD (Marchler-Bauer et al., 2015) are among the most popular protein annotation tools. Pfam is a comprehensive protein family database containing 16,712 families in the version 31. Each family is manually curated and is represented by a seed alignment and a Hidden Markov Model profile. A full alignment and the species distribution of all members collected by searching with the HMM profile against a sequence database are also provided. Pfam also groups homologous families into clans, if they cannot be merged (merged model fails to detect some sequences found individually).

The Conserved Domain Database (CDD) is a compilation of multiple sequence alignments from several other domain and full-length protein databases and domains contributed by NCBI projects (Tatusov, Koonin, & Lipman, 1997). The alignments are represented by position-specific score matrices (PSSM) and can be searched by the CD-search service at NCBI (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) (Marchler-

Bauer & Bryant, 2004). NCBI curated families in CDD consider structural information to delineate the alignment boundary and annotate functional sites when available. These families are also grouped into superfamily with a fine family hierarchy which is usually more function relevant.

When sequence-only methods fail to reveal more distant evolutionary relationship, spatial structures allow us to see further back in time, as protein structure is generally better preserved than sequence in evolution (Holm & Sander, 1996). Currently, the two leading structure classifications are SCOP (Structural Classification of Proteins) (Murzin, Brenner, Hubbard, & Chothia, 1995) and CATH (Class, Architecture, Topology, Homology) (Orengo et al., 1997), both of which are widely used in analyzing protein sequence, structure, function, and evolution and in developing various bioinformatics tools. CATH (http://www.cathdb.info) is largely automatic with added manual curation and emphasizes more on geometry, while SCOP is mainly manual and focuses on function and evolution.

In the SCOP (Murzin et al., 1995) (http://scop.mrc-lmb.cam.ac.uk/scop/index.html) hierarchical classification, closely related domains are grouped into families; families with structural and/or functional similarities supporting common ancestry are grouped into superfamilies; superfamilies with similar three-dimentional architectures and topologies are grouped into folds; and folds with similar secondary structure compositions are grouped into classes. Cataloging remote homologies identified by a combination of visual inspection, sequence and structure similarity search, and expert knowledge, the SCOP superfamily is the broadest level indicating homology and offers invaluable insights in protein evolution. However, SCOP tends to be conservative in assessing evolutionary relationships, and many homologous links reported in literature are not currently reflected (Aravind, Anantharaman, Balaji, Babu, & Iyer, 2005; Burroughs, Allen, Dunaway-Mariano, & Aravind, 2006; Burroughs, Balaji, Iyer, & Aravind, 2007; Copley & Bork, 2000; Nagano, Orengo, & Thornton, 2002; Ponting & Russell, 2000). Also, the recent dramatic increase of available structures in the PDB (Berman et al., 2000) (http://www.pdb.org) hinders careful manual curation in SCOP. Recently, a new version of SCOP (SCOP2) (Andreeva, Howorth, Chothia, Kulesha, & Murzin, 2014) was introduced that eschews hierarchical classification in place of a network of relationships (homologous and structural), although this database has not been made current with PDB. To partially alleviate this problem, ASTRAL now offers SCOPe, a sequence-based extension of the original SCOP hierarchy (Fox, Brenner, & Chandonia, 2014).

CATH is a hierarchical structure domain classification with four major levels: Class, Architecture, Topology, Homology. Class and Architecture describe the general secondary structure composition and shape at different degrees. Topology is similar to the fold in SCOP and considers secondary structures and their connectivity. Homology superfamily groups homologous domains together, with domains clustered at multiple redundancy and further divided into functional families (Sillitoe et al., 2015). Due to the large proportion of automation in its pipeline, CATH can process much more PDB structures than SCOP, but it is still far from complete.

Here I describe my contribution to the project in the lab to develop an evolutionary classification database of protein domains (ECOD), which catalogs much more distant homology relationships than others. Then I discuss how I improve the classification of family level in ECOD and create a set of family alignments and HMM profiles based on structural information from ECOD domains.

CHAPTER TWO

Evolutionary classification of protein domains with spatial structures

Introduction

The billions of proteins in extant species constitute a bewilderingly diverse protein world. To understand this world, systematic classifications are needed to reduce its complexity and to bring order to its relationships. As proteins are the products of evolution, their phylogeny provides a natural foundation for a meaningful hierarchical classification. As in the classification of species, a phylogenetic classification of proteins identifies evolutionary relationships between proteins and groups homologs (proteins that are descendants of a common ancestor) together. Because homologs generally share similar three-dimensional structures and functional properties, such a classification provides a valuable platform for studying the laws of protein evolution by comparative analysis as well as for predicting structure and function by homology-based inference.

Currently, the most widely used protein structure classifications are SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/index.html) (Murzin et al., 1995) and CATH (http://www.cathdb.info/) (Orengo et al., 1997). CATH is more reliant on automatic methods for classification, whereas SCOP relies more on manual analysis and curation. SCOP and CATH are invaluable resources for studying proteins, but they do not generally include the most recently solved structures in the PDB (Berman et al., 2000). Our classification ECOD maintains that the most recently determined structures, especially those evolutionarily distant from classified proteins, attract the most interest and hence are the most important to classify quickly and accurately.

Here I introduce the ECOD (Evolutionary Classification Of protein Domains) database. Our goal is threefold: (1) to construct a comprehensive domain classification based on evolutionary connections, (2) to extend the realm of connections to include remote homology, and (3) to maintain concurrent updates with the PDB. Because experimental data is very sparse compared to sequence data, establishing an evolutionary-based classification scheme of structures allows for biological insight into related proteins that otherwise lack functional information. In such a scheme, close homologs admittedly represent the most relevant source of functional inference. However, for most proteins, only distant homologs have been studied in detail. Fortunately, many examples have shown that analysis of proteins in the context of their distant homologs provides functional clues that advance biological research (Bazan & de Sauvage, 2009; Chai et al., 2003; Coles et al., 2005; Grishin, 2001c). In addition, remote homology offers deeper insights in protein evolution. In order to extend distant evolutionary relationships beyond the SCOP superfamily level in ECOD, I and other members in ECOD team apply state of the art homology-inference algorithms both developed in our group (Cheng et al., 2008; B. H. Kim, Cheng, & Grishin, 2009) as well as by others (Holm & Sander, 1993; Soding, 2005), manually analyze and verify the suggested homologous links, and incorporate findings from literature. For weekly updates, I rely on a computational pipeline that automatically and confidently classifies the majority of newly released structures and flags incompletely classified and unclassifiable structures, as well as a web interface that presents those difficult to deal with structures and pre-computed data in a convenient way for rapid manual inspection and classification.

ECOD is a publicly available database (http://prodata.swmed.edu/ecod/). By focusing on remote homology and weekly updates, ECOD strives to provide a more simplified and up-to-date view of the protein world than is currently available in existing classifications. As such, ECOD is unique in combining the following features: 1) the aforementioned weekly updates, following new releases from the PDB; 2) a hierarchy that specifically incorporates sequence-based relationships in a family level of close homology; 3) a classification that reflects more distant evolutionary connections; 4) a hierarchy that lacks a SCOP-like fold level, as the definition of “fold” is often subjective (Hadley & Jones, 1999); 5) domain partitions for all former members of the SCOP multi-domain protein class; and 6) combination of membrane proteins with their soluble homologs where an evolutionary relationship can be hypothesized. Theoretically, ECOD catalogs rich and up-to-date information about protein structure for the studies on protein origins and evolution; and practically, it helps homology-based structure and function prediction and protein annotation by providing a pre-compiled search database.

Database description

ECOD is a hierarchical classification of domains based on their evolutionary relationships. Focusing on remote homology, ECOD organizes domains into very broad homologous groups. At the same time, ECOD families address closer evolutionary relationships, detectable at a sequence level. Most importantly, ECOD is comprehensive and up-to-date, including all entries in the PDB and updating weekly, thus uniquely providing researchers with the most current classification of protein domains at both distant and close homology levels.

ECOD is a hierarchical classification with five main levels (Fig. 2.1, from top to bottom): architecture (A), possible homology (X), homology (H), topology (T), and family (F). The architecture level (A) groups domains with similar secondary structure compositions and geometric shapes. The possible homology level (X) groups domains where some evidence exists to demonstrate homology (but where further evidence is needed). The homology level (H) groups together domains with common ancestry as suggested by high sequence-structure scores, functional similarity, shared unusual features (Murzin, 1998), and literature. The topology level (T) groups domains with similar topological connections. The family level (F) groups domains with significant sequence similarity.

ECOD has 20 architectures that were developed both by consulting SCOP fold descriptions and inspecting numerous structures. It is worth noting that clear-cut boundaries between architectures do not always exist and that domain assignment to an architecture is sometimes subjective. This level is introduced largely for convenience of users and does not directly correspond to evolutionary grouping. A-level lies in between SCOP class and fold and groups proteins by simple visual features such as bundles, barrels, meanders, and sandwiches. Coiled-coils, peptides, fragments, largely disordered structures, and low-resolution structures were put in special architectures with no X-, H-, T-, or F-levels, as confident evolutionary classification of these structures is challenging at the moment. Nucleic acids, in addition to proteins, are kept within a special architecture and are not currently classified. Within architectures, X-groups are ordered by structural similarity between them.

The ECOD X-level groups domains that may be homologous as is frequently suggested by similarity of their spatial structures. A domain's overall structure is traditionally referred to as its ‘fold’. Fold similarity usually refers to general resemblance in both architecture and topology and can result from either common ancestry (homology) or physical/chemical restrictions (analogy) (Finkelstein & Ptitsyn, 1987; Krishna & Grishin, 2004; Orengo, Sillitoe, Reeves, & Pearl, 2001). Both SCOP and CATH have a fold level in the hierarchy: “SCOP fold” and “CATH topology”. However, the definition of fold can be subjective (Hadley & Jones, 1999), and fold is a geometrical concept without explicit evolutionary meaning. Therefore, ECOD generally avoids the fold concept. However, domains that share strong overall architectural and topological similarity and are possibly homologous, but which lack further evidence to exclude analogy, are attributed to the same X-group but different H-groups. The conceptual difference between ECOD X-group and SCOP fold can be shown, for example, in the classification of domains with a ferredoxin-like topology. In SCOP, the ‘Ferredoxin-like’ fold is a large assembly of various superfamilies that share the (βαβ)×2 topology. Among all these superfamilies, 4Fe-4S ferredoxins seem unique for their small size and cysteine-rich nature (cysteines are used to coordinate the Fe-S clusters). Thus, I suspect 4Fe-4S ferredoxins have an independent evolutionary origin and keep 4Fe-4S ferredoxins and other superfamilies in separate X-groups. On the other hand, although domains in the SCOP fold ‘Ribosomal proteins S24e, L23 and L15e’ do not have the ferredoxin-like (βαβ)×2 topology, their structures can easily be transformed into that topology by a circular permutation. Their structural similarity and functional similarity with the ‘RNA-binding domain, RBD’ superfamily in SCOP ‘Ferredoxin-like’ fold may imply homology. Therefore, ECOD classifies ‘Ribosomal proteins S24e, L23 and L15e’ and ‘RNA-binding domain, RBD’ as two H-groups in the same X-group as possible homologs. When further evidence coming either from additional sequences or three-dimentional structures accumulates, classification decisions are adjusted to agree best with all available data.

An ECOD H-group can contain more distant homologous links than the equivalent SCOP superfamily or CATH homologous superfamily. Although the majority of ECOD H-groups contain only a single SCOP superfamily (88%) or CATH homologous superfamily (81%), some H-groups contain many more. For example, the Immunoglobulin-related and the Rossmann-related H-groups contain the most SCOP superfamiles (47 and 28, respectively) and CATH homologous superfamilies (81 and 40, respectively). Superfamilies were merged based on multiple high-scoring homologous links between domains. These merges reflect the homology between domain members of these previously split groups.

To readily incorporate the observation that homologs can adopt different folds, ECOD has a topology (T-) level below the homology (H-) level. As a result, homologs with different topologies that SCOP necessarily separates into different folds (and thus different superfamilies) are unified in the same H-group but different T-groups in ECOD. For example, β-propellers are comprised of differing numbers of repeated β-meanders, all of which are evolutionarily related. The five different beta-propeller folds outlined in SCOP are organized in ECOD into a single H-group, with child T-groups for domains with differing number of blades (Chaudhuri, Soding, & Lupas, 2008). Also, the domain contents of 11

SCOP folds are organized into multiple T-groups under the Rift-related H-group in the cradle-loop barrel X-group (Alva, Koretke, Coles, & Lupas, 2008). If sufficient evidence for homology between these proteins is found, this consideration results in merging not only SCOP superfamilies, but also SCOP folds.

Within T-groups, ECOD organizes domains into families based on sequence similarity. I employ Pfam as the standard for family definition. ECOD domains were attributed to Pfam families by HMMER3 (Mistry, Finn, Eddy, Bateman, & Punta, 2013). Therefore, the majority of ECOD F-groups are simply Pfam families. However, not all protein domains with known structure can be attributed to even the latest version of Pfam by sequence similarity. Initially, those domains were grouped into families by HHsearch as provisional families. Then I endeavor to create an improved set of families based on ECOD structural domains, which is described in detail in Chapter 4.

Summary statistics of ECOD and comparison with SCOP and CATH

Summary statistics for the ECOD database as of October 27th, 2017 (version 199) are presented in Table 1. The majority of the 575,133 domains in ECOD were assigned automatically to a smaller set of 21,206 manually curated domain representatives. ECOD provides for domains which are assembled from multiple PDB chains, either due to photolytic cleavage (i.e. order-dependent assembly) or obligate multimers (i.e. order-independent assemblies). For order-independent assemblies, I distinguish between those domains where the assembly is primarily relevant for display, or appears to be biologically necessary. These are fairly rare in the database.

The growth of the PDB over time was compared to the number of structures classified in ECOD, CATH, and SCOP (Fig. 2.2(a)), data provided by Dr. R. Dustin Schaeffer. The difference between the number of structures in the PDB and those in the main architectures of ECOD can be primarily accounted for by the number of structures contained in ECOD special architectures (i.e. coiled-coil, peptide, non-peptide polymers, and low-resolution structures that could not be classified by sequence). The growth of the hierarchical levels from 2000–2013 indicates that although evolutionary distinct groups (i.e. X- and H-groups) are being discovered at a steady pace, the predominant source of new domains in ECOD is from sequence families (F-groups) being associated with existing homologous groups (Fig. 2.2(b)).

I analyzed the distribution of domains in hierarchical levels in ECOD. The most populated homologous groups (H-groups) are placed in context with their architecture in ECOD (Fig. 2.3(a)) and are also ranked by population (Fig. 2.3(b)). The Ig-related and Rossmann-related H-groups, in addition to containing the most merged SCOP and CATH homologous groups, are the most populated H-groups in ECOD. The merging of many previously distinct helix-turn-helix (HTH) SCOP superfamilies in ECOD boosts the population of this H-group considerably compared to its original SCOP population. The inset (Fig. 2.3(b)) shows those most populated H-groups by number of F-groups. Where many sequence families have been merged by distant homology, such as the RIFT-related or Immunoglobulin-related domains, H-groups will contain many F-groups. In ECOD, as opposed to SCOP or CATH, there exist fewer distinct homologous groups with related topologies, as many of these groups have been linked by homology. For example, in ECOD, there is a single Rossmann-related H-group among the most populated (top 15) groups, whereas in the most populated SCOP superfamilies or CATH homologous superfamilies, there are two (NAD(P)-binding Rossmann fold domains and SAM methyltransferases) and four (3.40.50.720, 3.40.50.1820, 3.40.50.150, and 3.40.50.2300), respectively.

I compared ECOD H-groups to SCOP superfamilies and folds by considering sequence and structure similarity of domain pairs within each level. ECOD manual representatives and ASTRAL40 domains were evaluated by HHsearch to reflect sequence similarity and TM-align to reflect structure similarity (Soding, 2005; Zhang & Skolnick, 2005). SCOP superfamilies tend to contain more close homologs that can be detected by sequence homology search methods than ECOD H-groups (Fig. 2.4(c)). Domains classified in SCOP folds (excluding pairs from the same superfamily) emphasize structural similarity, as the distribution is mostly populated in the low sequence similarity region and the peak shifts right compared with others (Fig. 2.4(a,b)). On the other hand, as ECOD H-group readily incorporates homologous links from SCOP superfamilies and also many remotely homologous relationships that were previously overlooked, its peak sizes lie between SCOP fold and superfamily in high and low sequence similarity regions. Also it is worth noting that the peak of ECOD H-group does not have the right shoulder in the intermediate sequence similarity group but has a relatively evident left shoulder in the high sequence similarity group (Fig. 2.4(b,c)), which potentially supports the idea that ECOD classification is homology-centric.

Automatic and manual classification in ECOD

A pilot version of ECOD based on SCOP 1.75 (Murzin et al., 1995) was first developed in lab by Dr. Hua Cheng and *et al*. To detect remote homologies beyond the SCOP superfamily level, 40% identity domain representatives in the first 7 classes in SCOP 1.75 were retrieved from ASTRAL (Chandonia et al., 2004) and compared in an all-versus-all fashion. Four scores were computed for each pair: HHsearch probability (Soding, 2005), DALI Z-score (Holm & Sander, 1993), HorA combined score (B. H. Kim et al., 2009), and HorA SVM score (Cheng et al., 2008). Domain pairs with high scores were manually inspected and analyzed. The decision on whether any given pair is homologous was based on considerations of the aforementioned scores, literature, functional similarity (such as common cofactor-binding residues), shared unusual structural features (Murzin, 1998), domain organization, oligomerization states, and disulfide bond positions. Since the SCOP superfamily level is reliable and conservative, typically only SCOP superfamilies were merged into homologous (H-) groups. In addition, Hua split SCOP entries with multiple domains or with duplications, and corrected rare inconsistencies in the SCOP classification. Cytoscape (Shannon et al., 2003) clustering was used to aid manual analysis by displaying domains and high-scoring links. After 40% representatives were classified, other SCOP 1.75 domains were automatically mapped into the ECOD hierarchy using MUSCLE alignments (Edgar, 2004). Many hierarchical groups in the ECOD pilot version retained the names of their original SCOP counterparts.

Those structures not classified in SCOP 1.75 and new structures in updates were partitioned and assigned to ECOD using a pipeline of a combination of sequence and structural homology detection methods, mainly developed by Dr. R. Dustin Schaeffer, which includes three sequence homology detection methods of increasing sensitivity and decreasing specificity to partition input proteins into domains. First, the input protein sequence is queried against a library of known ECOD full-length chains (containing both single-domain and multi-domain architectures) using BLAST (Camacho et al., 2009). Where significant sequence similarity (E-value<2e-3) is detected to a known domain architecture with high coverage (<10 residues uncovered), the entire series of domains in the input chain was partitioned in one pass. Second, the protein sequence is queried using BLAST against a library of domain sequences. Here single-domain proteins and components of multi-domain proteins were assigned individually by sequence similarity (E-value<2e-3) and hit coverage (>80%). Finally, for detection of more distant homology, a query sequence profile was generated using HHblits (Remmert, Biegert, Hauser, & Soding, 2011). This profile was used to query a database of ECOD representative domain profiles using HHsearch. Domains from the input chains could be classified by any combination of the three sequence-based methods (chain BLAST, domain BLAST, or domain HHsearch). Following partition, a boundary optimization procedure based on the structural domain parser, PDP, was run to eliminate small interstitial gaps between assigned domains and at termini (Alexandrov & Shindyalov, 2003). If a protein chain could not be assigned by the sequence pipeline, it was queried against a library of representative ECOD domain structures using DaliLite (Holm & Park, 2000). Domains were assigned where significant structural similarity existed to a known ECOD domain and where the aligned region passed a simple BLOSUM-based alignment score (Cheek, Qi, Krishna, Kinch, & Grishin, 2004).

Proteins that cannot be classified confidently and completely by automated methods are manually curated. The manual classification process involves partitioning the query protein into domains and identifying homologs or possible homologs for each domain (Cheng, Liao, Schaeffer, & Grishin, 2015). In this process, I rely on scientific literature, sequence and structure similarity comparison programs, popular protein databases (e.g., Pfam, SCOP, and CATH), visual inspection and comparison, as well as my knowledge and experience. I first inspect the mapping suggested by the pipeline. Oftentimes, the suggested mapping is correct for most or part of the query structure, and I typically accept this mapping but modify the domain boundaries. When a homologous hit with similar topology can be found, the query is classified into the same T-group as the hit; when a homologous hit with different topology can be found, the query is classified in a new T-group but the same H-group as the hit; when only a possibly homologous hit with similar overall structure can be found, the query is classified in a new H-group but the same X-group as the hit; when no possible homologs can be identified, the query is classified in a new X-group by itself. For proteins sharing remote homology, multiple factors usually need to be taken into consideration. A detailed example is illustrated in Chapter 3, where I elaborate on the discovery of homologous relationship between GAIN, ZU5 and NUP98-C domains.

Here I describe an example showing homology relationship with undetectable sequence similarity but pronounced structure similarity. Detection of structural similarity is often necessary for identifying evolutionary relationships between distant homologs. The mitochondrial RNA binding protein complex consists of two homologous proteins, MRP1 and MRP2, which bind to guide RNAs and are essential for kinetoplastid RNA editing in trypanosomatids (Aphasizhev, Aphasizheva, Nelson, & Simpson, 2003; Vondruskova et al., 2005). Although sequence homology detection methods, such as PSI-BLAST (Altschul et al., 1997) and HHsearch (Soding, 2005), fail to detect any other homolog of MRP, the crystal structures of MRP1 and MRP2 exhibit remarkable structural similarity to the Whirly family of single-stranded DNA (ssDNA) binding proteins in plant (Fig. 2.5(a,b)) (Schumacher, Karamooz, Zikova, Trantirek, & Lukes, 2006). A Dali alignment between MRP1 (PDB 2GIA) and WHY1 (PDB 1L3A) has a Z-score of 13.8 and a RMSD of 2.4 Å over 125 residues (Fig. 2.5(c)). I note that the structure prediction server I-TASSER (Roy, Kucukural, & Zhang, 2010) identified WHY2 (PDB 3N1H, ranked 2nd) and WHY1 (PDB 4KOO, ranked 9th) with normalized Z-scores of 0.73 and 0.53 respectively in the top 10 templates used for threading when MRP2 sequence is provided as input and close templates in PDB sequences are excluded. Whirly proteins bind to ssDNA functioning in transcription regulation and DNA double-strand break repair (Cappadocia et al., 2010; Desveaux, Allard, Brisson, & Sygusch, 2002; Desveaux et al., 2004) and also can bind to plastid RNA in chloroplast RNA metabolism (Krause et al., 2005; Prikryl, Watkins, Friso, van Wijk, & Barkan, 2008). In addition to the structural similarity of the protomer, MRP and Whirly proteins both form tetramers that superimpose well with a RMSD of 3.7 Å over 248 residues (Fig. 2.5(d)) (Desveaux et al., 2002; Schumacher et al., 2006). MRP complex is a heterotetramer with two MRP1 and two MRP2 (Schumacher et al., 2006), while Whirly proteins form a homotetramer (Cappadocia et al., 2010; Desveaux et al., 2002), and are suggested to further assemble into a 24-mer (Cappadocia et al., 2012). They also both bind to nucleic acids on the same surface in a sequence-independent fashion (Fig. 2.5(d)). However, distinct binding mechanisms are adopted. For MRP, binding is dominated by the electrostatic interaction between the positively charged surface of MRP and the phosphate groups of the guide RNA (Schumacher et al., 2006). Whirly proteins mainly use hydrophobic interactions of nucleobases and the compensation of few sequence-specific interactions is observed in structures with different ssDNAs (Cappadocia et al., 2010). The MRP and Whirly families represent two highly diverged branches that are distributed in animals plus trypanosomatids and plants, respectively. They likely originated from a duplication event as shown by other homologous families such as human transcription cofactor PC450, which is a homodimer of two ββββα units, and Pur-α whose bacterial homodimer structure (PDB 3N8B) (Graebsch, Roche, Kostrewa, Soding, & Niessing, 2010) and the duplicated form in Drosophila (PDB 3K44) (Graebsch, Roche, & Niessing, 2009) are solved. The results of this divergent evolution are reflected in the distinct sequence profiles of MRP and Whirly (Fig. 2.5(c)), posing a difficult challenge for sequence homology detection methods.

To facilitate manual analysis, I developed a web interface that presents pre-calculated data from the automatic pipeline, such as sequence and structural search scores, in tabular format together with an interactive molecular structure viewer, which allows coloring a hit or custom region on the query structure to get a better and intuitive idea of the domain boundary. The interface can record decisions and assignments from manual curators to the backend server, and the results will be automatically incorporated in ECOD during next update. I also added some utilities specifically for common problems seen in curation, for example, fast boundary adjustment to handle cases like just extending the domain boundary to the end of the chain.

There are still minor aspects in the pipeline and update automation that could be improved. Now the input to the classification pipeline is always a chain in PDB, so multiple chain domain and domain assembly are handled properly. In automatic domain assignment, the pipeline takes non-overlapping hits ranked by scores at each step independently and the accepted hits are passed on to the next step. It assumes that preceding method in the pipeline has better specificity and hit with higher score has better alignment, which works well for most cases but not all. For the best performance, I have suggested using dynamic programming to find the set of hits with largest coverage. For historic reason, the master manual classification was stored in a plain text file and later uploaded to a shared Google Docs document. It is independent of other information stored in our SQL server in lab and it is not a structured document. Therefore, there is currently no way for a program to modify it, which arguably is a bad thing or not. However, a structured database storing the assignment result from sources like the curation interface is needed for any kind of large scale or quick curation other than regular weekly updates. As the result, multiple records of manual classification exist at this moment. It would be the best if they can be unified and converted into a SQL database and a web interface is developed to allow input and modification from everyone at ease.

ECOD website

I developed the ECOD website and designed the SQL database that supports the website and enables fast access to various information in ECOD. The index ECOD webpage provides quick links to search, downloadable files, and an online browsable version of the hierarchy. Brief summary statistics are presented showing the number of PDBs and domains classified by the current version of ECOD. Users can search their interested protein using a FASTA sequence with BLAST (Camacho et al., 2009) against full ECOD domain database, or a structure in PDB format with TM-align (Zhang & Skolnick, 2005) against curated ECOD manual domains.

The search result page is divided into three sections: first, all ECOD domain hits are shown in a schematic overview that shows which regions of the query protein share similarity. The query protein is represented as a large grey bar at the top, and hit domains follow below and are colored by scores. Clicking on an individual result will navigate the user to the specific alignment results; Second, individual domain results with their ECOD hierarchal information and score statistics are shown following the schematic overview. The full names of the X-, H-, T-, and F-groups for each ECOD hit domain are shown. Finally, clicking on the ECOD domain id for any hit will navigate to that domain’s description page. Lastly, the alignments for individual hits along with their statistics and links to their domain description pages are presented in the final section of results. For TM-align results, links to an online JSmol (Hanson, Prilusky, Renjian, Nakane, & Sussman, 2013)viewer of the hit domain and a downloadable PyMol (Schrodinger, 2015) session containing the superposition between the query protein and hit domain are available.

The ECOD domain summary page consists of four main components. The summary page is titled with the ECOD domain identifier. Directly below are the five ECOD hierarchal levels into which the domain has been classified. In each case, clicking the magnifying glass icon next to the name of domain identifier or hierarchal level will navigate you to the position of that level in the ECOD tree view. The “Download files” dialog is available for the user to download FASTA format sequence files, PDB format structure files, or a pre-generated PyMol structure viewer session for the domain. The PyMol session can be generated containing the domain, the domain in the contest of its PDB chain, or the domain in the context of the entire PDB deposition. Images for the domain in these contexts are provided adjacent to the download dialog. Below the hierarchal levels of a domain, several domain characteristics are presented, including the unique numerical identifier and domain type showing whether the domain is manually curated or automatically assigned. Automatic domains will have a link to the manually curated domain with which they are associated in the “Parent” entry. If a domain was assigned by distant homology to a domain in another F-group, and nucleates the classification of a new F-group, that domain will be designated as a provisional manual representative. Non-peptide ligands within 4 Å of the domain are displayed in images and listed in the “Ligand” entry. Finally links to the PDB deposition and the title of the PDB chain are displayed. A JSmol structure viewer is placed in the “Structure View” tab. Different display modes are supported and display of ligands can be toggled. The “Domain Organization” tab provides brief descriptions of and links to domains sharing the same PDB structure and chain.

ECOD can be viewed as a tree with hierarchal levels and browsed interactively in the tree view page. Each node at each level can be expanded or collapsed as necessary. Manual representative domains are directly displayed beneath F-groups. All automatically assigned domains are associated with a manual representative, these domains can be viewed by clicking on the “nonrep” link. Sometimes provisional manual representatives are created from automatic domain if no manual representative was available within an F-group. These provisional domains are indicated by an asterisk. Links to PyMol session, image, and webpage JSmol viewer of domain are provided. Where available, PubMed and DOI links are provided for the primary citation in the PDB deposition. For non-singleton T-groups, a tree of component F-groups can be viewed by clicking the small tree icon on the F-group. Using the ECOD tree view, users can generate perspective on the known structure space near their domain of interest.

Figures and table

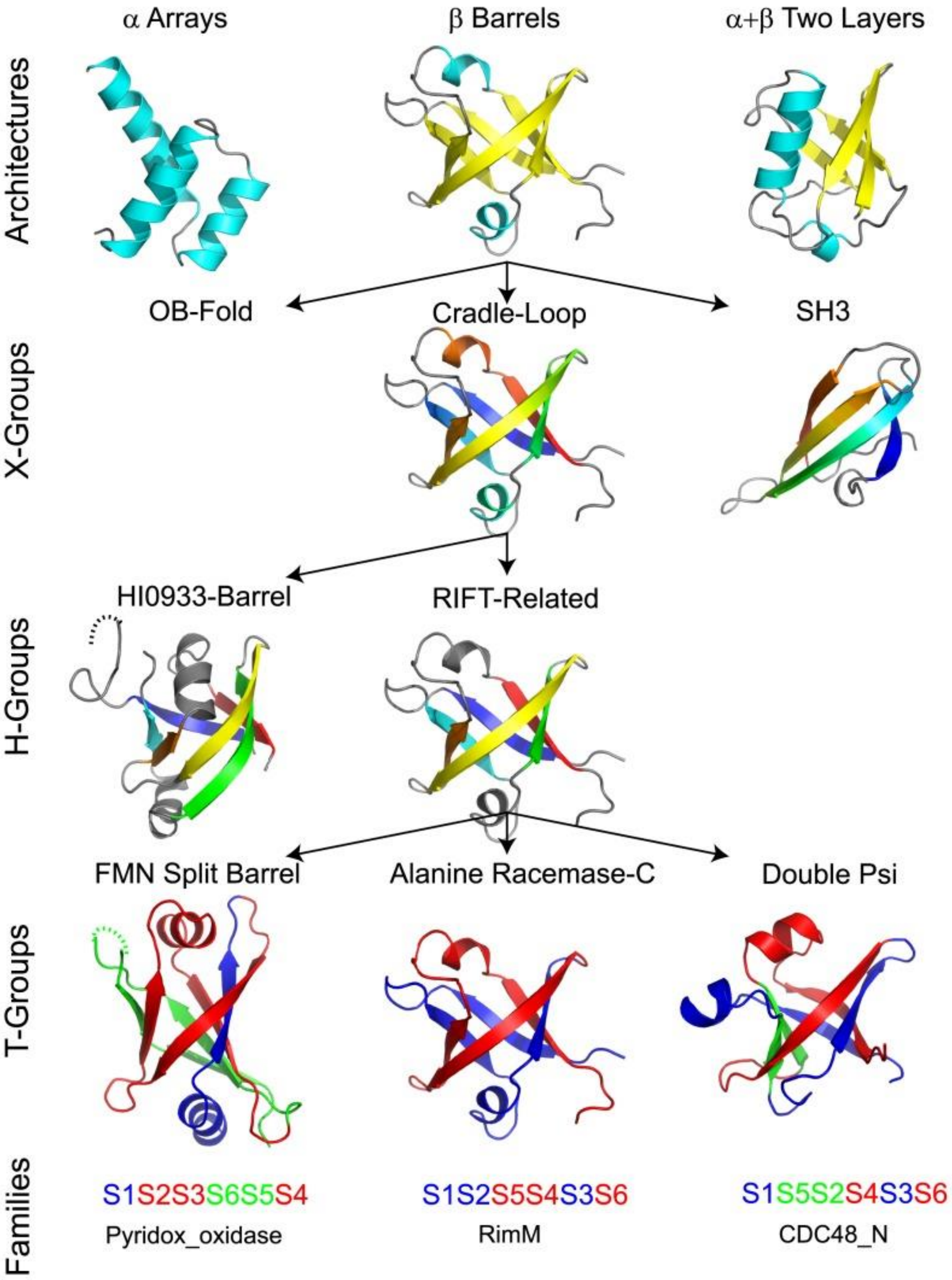


Fig. 2.1. Hierarchical levels of ECOD.

Domains placed within the same Architecture share similar secondary structure content (helix, cyan; sheet, yellow) and geometric arrangement. Domains placed within the same X-group share similar structure but lack a convincing argument for homology (vs. analogy), while those placed within the same H-groups are homologous. X- and H- group structures are colored in rainbow by consecutive secondary structure elements. T-groups distinguish homologous domains with notable differences in topology, such as the illustrated Rift-related metafold (Alva et al., 2008). Rift-related half-barrels (colored blue and red) are consistent among the domains, but permutations and strand swaps (green) modify the topology.

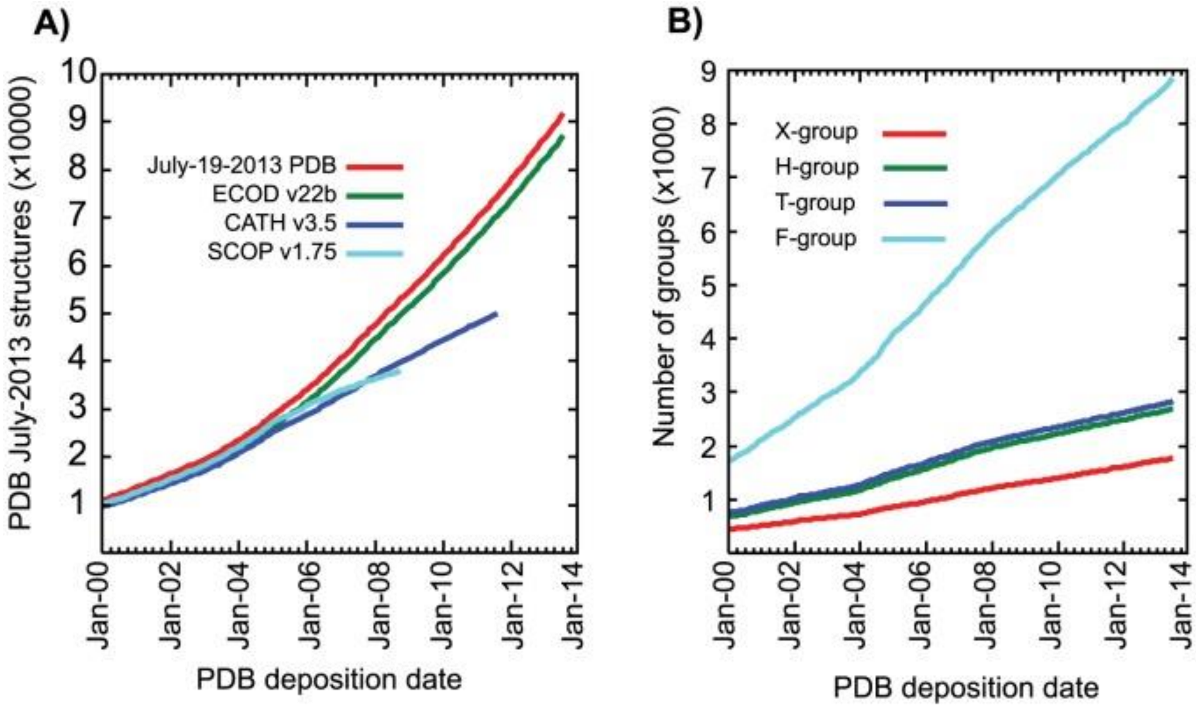


Fig. 2.2. Classification of ECOD and ECOD hierarchical levels with respect to the PDB and other classifications.

(a) A cumulative sum of PDB release dates from Jan-2000 to Jan-2014 (red) compared to classified PDB depositions in ECOD (green), SCOP (cyan), and CATH (blue). Any deposition with at least one domain classified is counted. ECOD consistently classifies more structures than SCOP and CATH and is more up-to-date. (b) The cumulative sum of PDB deposition dates in ECOD hierarchical levels. Each group is classified once by its oldest deposition. The number of new levels increases consistently over time over the 2000 to 2014 time period.

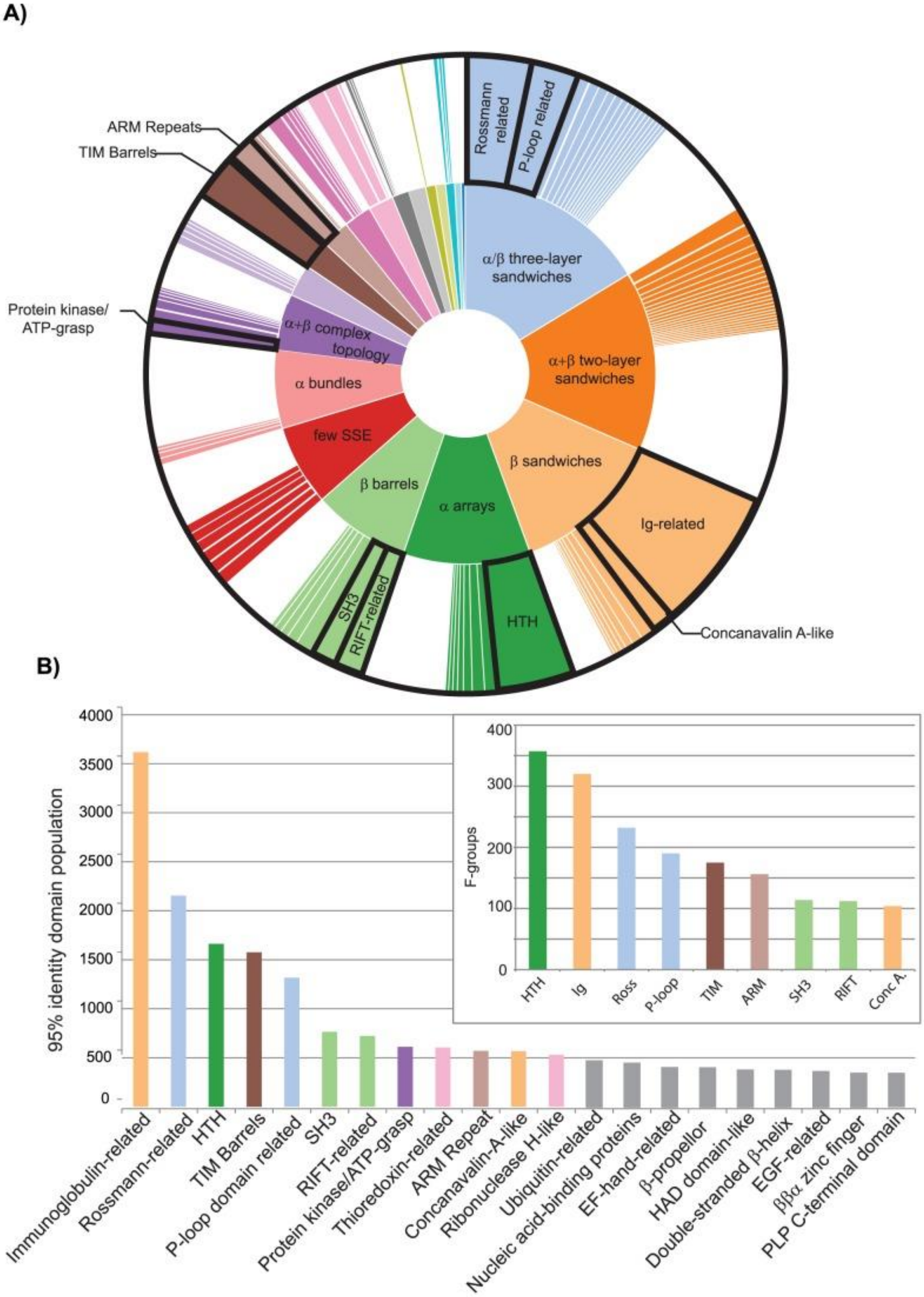

A)
ARM Repeats
TIM Barrels
Protein kinase/
ATP-grasp
Rossmann
related
P-loop related
α/β three-layer
sandwiches
α+β two-layer
sandwiches
β sandwiches
Ig-related
Concanavalin A-like
HTH
α arrays
β barrels
SH3
RIFT-related
few SSE
α bundles
α+β complex
topology
B)
95% identity domain population
4000
3500
3000
2500
2000
1500
1000
500
0
Immunoglobulin-related
Rossmann-related
HTH
TIM Barrels
P-loop domain related
SH3
RIFT-related
Protein kinase/ATP-grasp
Thioredoxin-related
ARM Repeat
Concanavalin-A-like
Ribonuclease H-like
Ubiquitin-related
Nucleic acid-binding proteins
EF-hand-related
β-propellor
HAD domain-like
Double-stranded β-helix
EGF-related
ββα zinc finger
PLP C-terminal domain
F-groups
400
300
200
100
0
HTH
Ig
Ross
P-loop
TIM
ARM
SH3
RIFT
Conc A.

Fig. 2.3. Distribution of H-groups in ECOD by architecture and 95% representative domain population.

(a) H-groups are colored by architecture and sized according to their representative domain population. H-groups smaller than 0.01 radians are not displayed. Those H-groups shown in bottom distributions are labeled. (b) The most populated H-groups (>500 95% representative domains) are colored by architecture. The immunoglobulin-related, Rossmann-related, and helix-turn-helix (HTH) H-groups are the most populated H-groups in ECOD. The inset shows the most populated H-groups by number of F-groups.

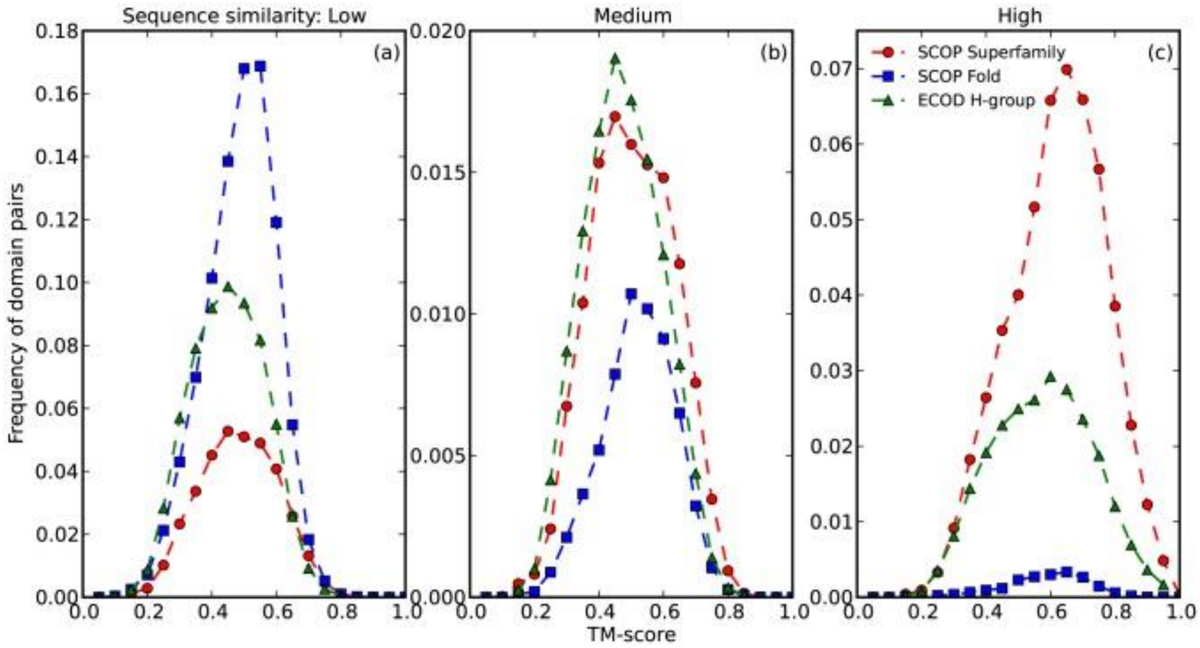


Fig. 2.4. Structure similarity distribution of domain pairs from SCOP superfamily, SCOP fold and ECOD H-group, measured by TM-score.

Data were grouped into three panels by sequence similarity in terms of HHsearch probability (Low: probability ≤20%, Medium: 20%<probability<90%, High: probability ≥90%) and then binned into 20 bins to calculate frequency.

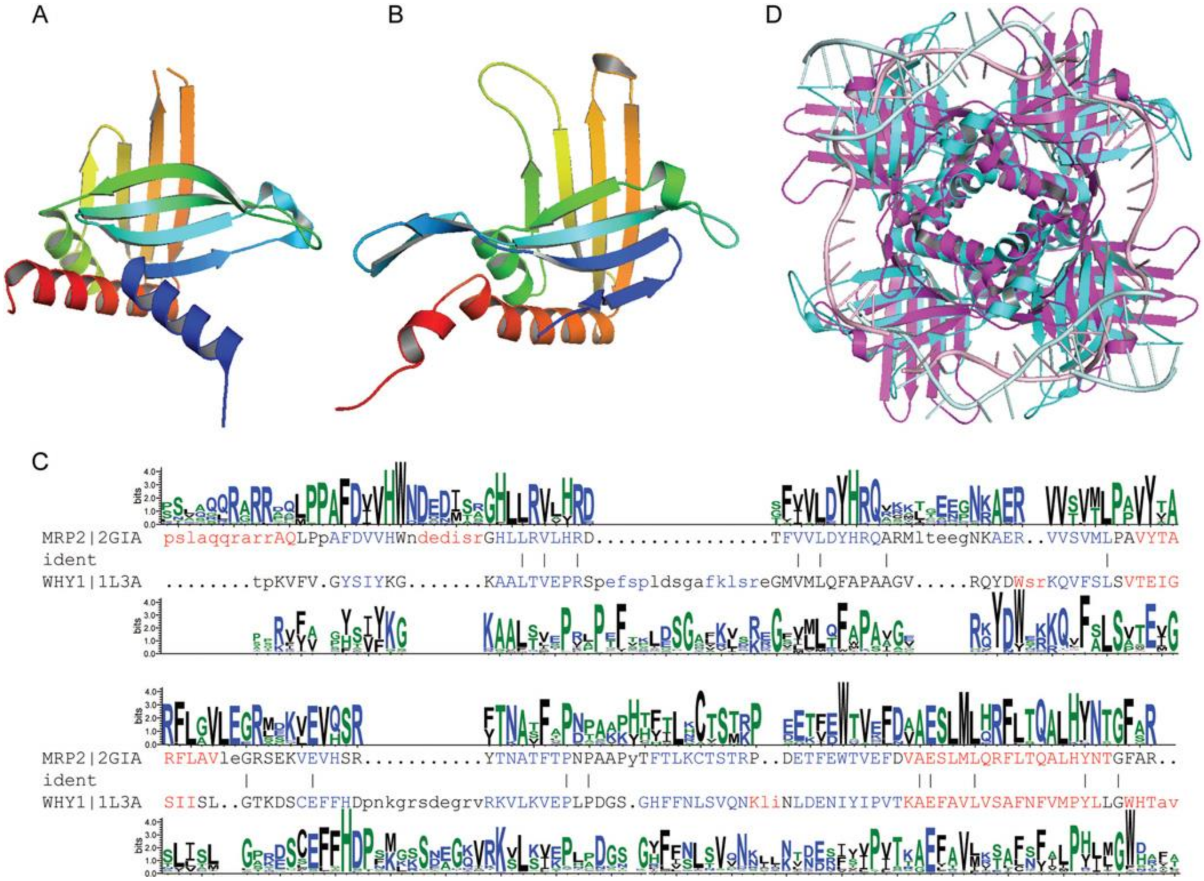

A
B
D
C
MRP2|2GIA pslaqqrarrAQLPpAFDVVHWndedisrGHLLRVLHRD................TFVVLDYHRQARMlteegNKAER..VVSVMLPAVYTA
ident
WHY1|1L3A ....tpKVFV.GYSIYKG.......KAALTVEPRSpefspldsgafklsreGMVMLQFAPAAGV.....RQYDWsrKQVFSLSVTEIG
MRP2|2GIA RFLAVleGRSEKVEVHSR...........YTNATFTPNPAAPyTFTLKCTSTRP..DETFEWTVEFDVAESLMLQRFLTQALHYNTGFAR..
ident
WHY1|1L3A SIISL..GTKDSCEFFHDpnkgrsdegrvRKVLKVEPLPDGS.GHFFNLSVQNKliNLDENIYIPVTKAEFAVLVSAFNFVMPYLLGWHTav
bits

Fig. 2.5. Structure and sequence comparisons of MRP and Whirly.

(a) Structure of MRP2 (PDB 2GIA, chain A). (b) Structure of WHY1 (PDB 1L3A, chain A). Both structures are shown in cartoon and colored in a rainbow. (c) Dali structure alignment of MRP2 and WHY1. Residues are colored red for α-helices and blue for β-strands. Sequence profiles are represented by sequence logos generated from multiple sequence alignment of BLAST hits by WebLogo (Crooks, Hon, Chandonia, & Brenner, 2004). (d) Superposition of MRP1/MRP2 (PDB 2GJE) and WHY2 (PDB 3N1K) tetramers with bound nucleic acids. MRP1 and MRP2 are colored cyan with RNA in pale cyan. WHY2 is colored magenta with DNA in light pink. Crystallography symmetry was applied to generate the biological units.

Table 2.1. Summary statistics of ECOD version 199 (Oct. 27, 2017)

LEVEL	POPULATION
Architectures	20
X-groups	2,232
H-groups	3,532
T-groups	3,739
F-groups	14,438
Manual representatives	21,206
Domains	575,133
PDB structures	134,111
Peptide chains	410,810

CHAPTER THREE

An ancient autoproteolytic domain found in GAIN, ZU5 and Nucleoporin98

Introduction

Proteolysis is a ubiquitous post-translational modification that can be as simple as removing an N-terminal methionine and signal peptide or activating precursor proteins to final mature products. But over the years, only a few protein families were found to contain domains with autoproteolytic activity, including early characterized Ntn-hydrolases, hedgehog proteins, inteins, pyruvoyl-dependent enzymes (Perler, Xu, & Paulus, 1997) and recently studied Nucleoporin98 (Nup98) (Hodel et al., 2002), SEA domain (Macao, Johansson, Hansson, & Hard, 2006), a DmpA/OAT superfamily hydrolase ThnT (Buller, Freeman, Wright, Schildbach, & Townsend, 2012) and GPCR-Autoproteolysis Inducing domain (Arac et al., 2012). A common activation mechanism of N-O or N-S acyl rearrangement is proposed (Perler et al., 1997). The activated serine, threonine or cysteine attacks the preceding peptide bond and results in an unstable ester intermediate, which then undergoes varying chemical reactions depending on the biological context.

The cell adhesion family of G protein-coupled receptors (GPCRs) is the second largest family of GPCRs in humans (Lagerstrom & Schioth, 2008). 33 adhesion GPCRs from 9 subfamilies have been discovered in human (Bjarnadottir et al., 2004) , but most of them remain to be orphan receptors (i.e. their endogenous ligands are unknown) (Langenhan, Aust, & Hamann, 2013). They feature long and diverse N-terminal extracellular domains and share a conserved autoproteolytic motif with polycystic kidney disease (PKD) proteins named

GPCR Proteolysis Site (GPS) (Langenhan et al., 2013). The GPS motif was discovered in latrophilin as a conserved sequence motif of about 40 amino acids with an autoproteolytic signature of HL↓T/S and always precedes the first transmembrane helix (Krasnoperov et al., 2002). Recently structures of fragments containing the GPS motif were solved for two adhesion GPCRs: latrophilin-1 and brain angiogenesis inhibitor 3 (BAI3) (Arac et al., 2012). Unexpectedly, these structures revealed that the GPS motif is an integral part of a larger beta-sandwich domain. Together with a preceding helix bundle subdomain, it is termed GPCR-Autoproteolysis Inducing (GAIN) domain, which is shown to be both necessary and sufficient for autoproteolysis (Arac et al., 2012). Subdomain A of GAIN domain (GAIN-A) is composed of six alpha-helices and the C-terminal subdomain B (GAIN-B) contains 13 beta-strands with the GPS motif covering the last 5 beta-strands.

Human Nup98 is encoded as a fusion of the Nup98 gene directly upstream of the Nup96 gene (Fontoura, Blobel, & Matunis, 1999). Both the Nup98-Nup96 precursor and the alternative splice variant Nup98 alone undergo autoproteolytic processing at the C-terminal domain of Nup98 with a conserved motif HF↓S (Rosenblum & Blobel, 1999). The removal of the short C-terminal fragment is required for Nup98 localizing to the nuclear pore, binding to Nup96 at the nuclear side of the nuclear pore complex (NPC) (Hodel et al., 2002) and also binding to Nup88 at the cytoplasmic side of NPC (Griffis, Xu, & Powers, 2003). Recently, Nup98 has also been found functioning as a transcription regulator. In *Drosophila*, Nup98 was shown to activate genes involved in development and cell cycle inside the nucleoplasm (Capelson et al., 2010; Kalverda, Pickersgill, Shloma, & Fornerod, 2010). In human cells, Nup98 interacts with different genome regions dynamically depending on the differentiation stage (Liang, Franks, Marchetto, Gage, & Hetzer, 2013). Overexpression of full-length Nup98 in neural progenitor cells leads to enhanced expression level of Nup98-associated neural developmental genes, but a fragment of Nup98 lacking the C-terminal domain decreased the expression level of those genes (Liang et al., 2013). Nup98 also fuses with many partner genes by chromosome translocation in patients with hematopoietic malignancies, resulting in chimeras with N-terminal FG repeats of Nup98 and C-terminal domains in other proteins, such as homeodomain, PHD zinc finger and coiled-coils (Gough, Slape, & Aplan, 2011). In yeast, there are three Nup98 homologs. The first, Nup145, is also cleaved autoproteolytically to give rise to two fragments, Nup145N and Nup145C, which are similar to Nup98 and Nup96 in human, respectively (Teixeira et al., 1997). The other two homologs, Nup116 and Nup100, only have the N-terminal part that corresponds to Nup98 and lack the autoproteolytic motif (Wente, Rout, & Blobel, 1992). The structures of the C-terminal domain in human Nup98, and yeast Nup145N and Nup116 have been determined experimentally (Hodel et al., 2002; Sampathkumar et al., 2010; Yoshida, Seo, Debler, Blobel, & Hoelz, 2011).

Initially discovered in zonula occluden (ZO)-1 and netrin receptor UNC5 (Leonardo et al., 1997), ZU5 domain manifests various functions in different proteins. The ZU5 domain in UNC5B binds to its death domain (DD) and prevents it from recruiting other components in apoptotic machinery by occupying the same interface for oligomerization (R. Wang et al., 2009). Disrupting this autoinhibition resulted in enhanced UNC5B activities in apoptosis and parachordal vessel formation in zebrafish (R. Wang et al., 2009). Two tandem ZU5 domains exist in ankyrins and the first ZU5 domain is solely responsible for binding to spectrin

(Ipsaro & Mondragon, 2010). Among all ZO proteins, only ZO-1 has an additional ZU5 domain at the C-terminus which is a minimal ZU5 domain with several strands missing (Huo et al., 2011). This ZU5 domain is thought to be responsible for interacting with cytoskeletal dynamics regulatory protein kinase MRCKβ and targeting it to the leading edge of migrating cells (Huo et al., 2011). Although no ZU5 domains with available structures are processed posttranslationally, p53-induced protein with a death domain (PIDD) and UNC5C-like protein with ZU5 domains are both cleaved by autoproteolysis constitutively both at HF↓S sites (Heinz et al., 2012; Tinel et al., 2007).

In this study, I demonstrate that GAIN, ZU5 and Nup98 C-terminal domain are distantly homologous and evolved from a common ancestor domain with autoproteolytic ability. Divergent families of bacterial, archaeal and eukaryotic homologs are identified, and human proteins are selected for discussion, providing insights of the evolution of these domains and their autoproteolytic motifs.

Homologous relationship of GAIN, ZU5 and Nup98 C-terminal domain

I studied the homologous relationships of the autoproteolytic GAIN domain to other domains. Interestingly, the C-terminal beta-sandwich subdomain B (GAIN-B) of BAI3 (PDB: 4DLO, residue range: 691 to 866) finds many ZU5 domains by the Dali Server (Holm & Rosenstrom, 2010) as top hits. The best hit (PDB: 3UD1, a ZU5 domain in human erythrocyte ankyrin) has Z-score 8.7, RMSD 3.1 Å and an alignment of 121 residues, whereas ZU5 domains are typically about 140 residues long. The ZU5 domain in UNC5B (PDB: 3G5B) can be aligned over 129 residues with Z-score 6.9 and RMSD 3.0 Å. 3UD1 is the reciprocal Dali best hit as it can also find 4DLO first (except for ZU5 domains). The structures of GAIN-B and ankyrin ZU5 domain both have a beta-sandwich fold that adopts the same topology and shares 11 beta-strands (Fig. 3.1(a,b)). They also share an unusual beta-hairpin (colored in wheat in Fig. 3.1), which is substituted by flexible loops in the minimal ZU5 domain in ZO-1 (Fig. 3.1(e,f)). When the sequence of BAI3 GAIN-B is submitted to HHpred (Soding, Biegert, & Lupas, 2005), a HMM-HMM based protein homology detection and structure prediction server, to search against the PDB database (September 6th 2014), it pulls out ZU5 domain in UNC5B (PDB: 3G5B) with probability 90.0%, E-value 4.2 and two other ZU5 domains in ankyrin-1 (PDB: 3F59) and ZO-1 (PDB: 2KXS) with slightly lower probabilities of 85.1% and 88.1%, and E-values 1.6 and 0.87, respectively.

ZU5 domains with available structures (ZO-1, UNC5B, ankyrin-1 and -2) and many ZU5 domains in human proteins lack the conserved serine or threonine and should not possess autoproteolytic capability, as shown in the multiple sequence alignment (Fig. 3.2). However, the experimentally verified autoproteolytic sites in PIDD (Tinel et al., 2007) and UNC5C-like protein (Heinz et al., 2012) are present in the equivalent positions with GAIN and Nup98 C-terminal domain (Fig. 3.2). Moreover, when mapped to other ZU5 structures, these motifs are also located at the corresponding positions in space. I suggest that some ZU5 domains lost autoproteolytic ability in evolution and developed divergent functions.

The structure of Nup98 C-terminal domain (Nup98-C) retains most of the core strands, although it is more structurally divergent with different insertions and deletions (Fig. 3.1(d)). Compared with ZU5 and GAIN domains, the N-terminal part varies significantly and misses the first beta-strand (colored in dark blue in Fig. 3.1); the insertion between beta-strands 4 and 6 (colored in green) is missing and the hairpin of beta-strands 8 and 9 is replaced by helices. Nevertheless, a Dali search with PDB 2Q5X (a cleavage-resistant Nup98 mutant) (Sun & Guo, 2008) as query finds the GAIN domain of BAI3 (PDB: 4DLO) with Z-score 4.7, RMSD 3.1 Å, alignment length 95 and ankyrin ZU5 domain (PDB: 3UD2) with Z-score 4.4, RMSD 3.7 Å and alignment length 99 immediately after top Nup98 hits. The Dali alignment covers the 7 beta-strands that constitute the core of the beta-sandwich. Although the structural scores are not compelling in and of themselves, an HHpred search against PDB database (September 6th 2014) using the sequence of ZU5 domain in ZO-1 also found several Nup98 homolog hits with moderate probabilities. The best Nup98 hit is PDB 3PBP chain B (yeast Nup116) with probability 86.3% and E-value 1.3. Searching using ZU5 domain in UNC5B can also find Nup98 with lower HHpred probability 61.0% and E-value 4. In return, when Nup98-C was used as query, ZU5 of UNC5B was detected as a hit with probability 60.3% and E-value 14. While these sequence alignments are relatively short, they cover beta-strands 7, 10 and 11 as labeled in Fig. 3.2 and finally extend to the autoproteolytic motif. More importantly, the HHpred sequence alignments are consistent with the Dali structure alignments, and the autoproteolytic site of Nup98 is at the equivalent position as in GAIN-B in both HHpred and Dali alignments. Taken together, I believe that GAIN, ZU5 and Nup98-C domains are remotely homologous based on structural and sequence scores, the consistency between Dali and HHpred alignments and their common autoproteolysis function.

Domain architectures and phylogenetic distribution of GAIN-B, ZU5 and Nup98-C homologs

I then used transitive PSI-BLAST to search for homologs starting from GAIN, ZU5, Nup98-C domain and remote homologs detected by HHpred. PSI-BLAST (Altschul et al., 1997) hits were first clustered by BLASTCLUST with score coverage threshold 0.5 (-S option). Representative sequences of each cluster were used to initiate new PSI-BLAST searches. Such a procedure was repeated until convergence. I first focused on proteins in the human proteome. I depict domain architectures of representative proteins from 9 adhesion GPCR subfamilies (Bjarnadottir et al., 2004), as well as PKD proteins in Fig. 3.3, and the domain architectures of other adhesion GPCRs can be viewed in reference (Lagerstrom & Schioth, 2008). The Pfam family DUF3497 (Pfam: PF12003), when mapped to the latrophilin structure, spans GAIN-A and the N-terminal part of GAIN-B before the GPS motif. Therefore, they are replaced by GAIN-A and GAIN-B in Fig. 3.3. Interestingly, ZU5 domains found in human are all intracellular while GAIN domains are extracellular and always located just before the first transmembrane helix (Fig. 3.3(a,b)). Collectively I refer to ZU5 domains in these proteins as canonical ZU5 domains. I observed that the consecutive ZU5, UPA and Death domain (DD) architecture is commonly used in all human proteins containing the canonical ZU5 domain except for ZO-1 (Fig. 3.3(b)). Such a domain organization was firstly discovered in UNC5B structure and thought to be shared by UNC5, PIDD and ankyrin from which the UPA domain got its name (R. Wang et al., 2009).

Of these 13 human proteins, UNC5 is a family of single-pass membrane proteins composed of the same ZU5-UPA-DD domain organization in the cytoplasmic component. UNC5A, UNC5B, UNC5C and UNC5D have similar extracellular domain architectures. In contrast, UNC5C-like protein has only a very short extracellular terminus and contains an autoproteolytic site of HFS motif in the ZU5 domain (Heinz et al., 2012). Others are cytoplasmic proteins and have some variation of the general ZU5-UPA-DD domain organization. SH3BP4 (SH3 domain-binding protein 4) (Y. M. Kim et al., 2012) and MACC1 (metastasis-associated in colon cancer 1) (Stein et al., 2009) contain two tandem DD domains and also an SH3 domain inserted before the DD. The ankyrin family in human consists of three members which have a ZU5-ZU5-UPA-DD domain arrangement in the central region. The two ZU5 domains in them are thought be functionally different (C. Wang, Yu, Ye, Wei, & Zhang, 2012). Compared with UNC5B, the first ZU5 domain of ankyrins interacts with the UPA domain in a similar fashion; the second ZU5 domain protrudes out with fewer interactions and the DD is not sequestered by any of the ZU5 domains (C. Wang et al., 2012). PIDD and the less studied death domain-containing protein 1 (DTHD1) (Abu-Safieh et al., 2013) also comprise a similar domain architecture of two consecutive ZU5 domains followed by one UPA and one DD, which may adopt a similar structure of that in ankyrins.

Two human proteins involved in the inflammasome, CARD8 (caspase recruitment domain-containing protein 8) and NLRP1 (NACHT, LRR and PYD domains-containing protein 1) contain a “Function to Find” (FIIND) domain (D'Osualdo et al., 2011). The FIIND domain is only present in chordates and cannot be linked to canonical ZU5 domains by PSI-BLAST. However, it also consists of a ZU5-like domain and a UPA-like domain according to D'Osualdo *et al*. (D'Osualdo et al., 2011). Here I refer to the ZU5-like domain in the FIIND family as ZU5-FIIND. As the CARD domain belongs to the DD superfamily (Park et al., 2007), CARD8 and NLRP1 also share a divergent ZU5-UPA-DD domain organization

(Fig. 3.3(c)). HHpred locates a high-scoring structure template for the middle region of NLRP1 in NLRC4 (PDB: 4KXF) (Hu et al., 2013), which contains a NACHT domain and a winged helix-turn-helix domain (Fig. 3.3(c)).

I also discovered a ZU5-like domain by transitive PSI-BLAST in two human protein families, cartilage intermediate layer protein (CILP) and the uncharacterized family FAM171. Full-length FAM171 proteins are annotated as the UPF0560 in Pfam (Finn et al., 2014), but this family possesses an extracellular ZU5-like domain before the predicted transmembrane helix (Fig. 3.3(d)). CILP proteins are secreted proteins in cartilage which are further cleaved into two chains (Bernardo et al., 2011; Lorenzo, Neame, Sommarin, & Heinegard, 1998). The N-terminal part of CILP-1 was determined to be an IGF-1 antagonist (Johnson, Farley, Hu, & Terkeltaub, 2003) and an inhibitor of TGF-beta1 signaling (Seki et al., 2005). Lorenzo *et al*. proposed that CILP-1 and CLIP-2 are processed upon secretion by a furin-like protease at a predicted consensus site RRNKR↓EDRT (Lorenzo et al., 1998). In our alignment, this site (Fig. 3.2, abbreviated as "(9)" in blue) is located one strand prior to where the common cleavage motif is located. These ZU5-like domains in CILP and FAM171 (together named ZU5-CF) are expressed in the extracellular space and represent a diverse branch of ZU5 domains with deteriorated autoproteolytic motifs (Fig. 3.2).

In addition to focusing on human proteins, I also discovered numerous ZU5 domain homologs in archaea (462 sequences) and bacteria (3,241 sequences) in our transitive PSI-BLAST sequence searches. For comparison, I also collected over 10,000 sequences of eukaryotic homologs. The sequence clusters of the complete set, when visualized by CLANS (Frickey & Lupas, 2004), reveal that canonical ZU5 domains and bacterial homologs (in red) cluster together, while clusters of eukaryotic remote homologs, such as GAIN, Nup98, and ZU5-FIIND domains (in green), scattered around the periphery of the diagram (Fig. 3.4). The Pfam family DUF1191 was another distantly homologous group identified by HHpred and is only found in green plants. Most proteins containing DUF1191 domains are single-domain proteins with a predicted signal peptide (Fig. 3.3(f)). The Pfam DUF1191 domain includes the region homologous to GAIN/ZU5/Nup98-C and a C-terminal predicted transmembrane helix. A group of archaeal homologs were also found (blue in Fig. 3.4), with the ZU5-like domain mapped to a previously defined PGF_pre_PGF domain in TIGRFAM (Haft et al., 2013). Many of these archaeal proteins are annotated as cell surface proteins and frequently contain domains involved in cell adhesion such as PKD and CARDB (cell adhesion related domain found in bacteria) domains (Fig. 3.3(g)).

As previously mentioned, extracellular ZU5-CF domains represent a divergent group which is demonstrated by two small close groups (one for CILPs and one for FAM171 proteins) clustered with some bacterial homologs separated from canonical ZU5 domains (Fig. 3.4). Bacterial ZU5 homologs are very diverse as shown by loose clusters spreading in the middle. The majority of these bacterial homologs possesses a predicted signal peptide and thus is likely exported from the cell. About half of them retain the conserved HFS motif, indicating that such autoproteolytic domains arose early in evolution and may have lost the autoproteolytic function and diverged to gain other functions. These bacterial proteins largely remain uncharacterized and their domain organizations vary among clusters. But the bacterial ZU5-like domain is most commonly observed to precede three or more SLH (S-layer homology) domains (A. Lupas et al., 1994) and sometimes together with the RHS/YD repeats which were recently revealed to form a large cocoon encapsulating the toxin (Fig. 3.3(h)) (Busby, Panjikar, Landsberg, Hurst, & Lott, 2013).

In agreement with the previous study (Arac et al., 2012), our search found sequences closely related to GAIN domain in a wide variety of eukaryotic branches, including Filozoa, Amoebozoa, Excavata and Alveolata. Some of these homologous sequences are clustered with metazoan sequences by BLASTCLUST and others form a small cluster by themselves. In a previous study (Krishnan, Almen, Fredriksson, & Schioth, 2012), adhesion GPCRs were suggested to emerge after the split of unikonts from bikonts. But GAIN domain was discovered in GPCRs in *Naegleria gruberi* (e.g. Genbank: XP_002674282.1), a bikont in the Excavata supergroup, suggesting a much earlier origin of adhesion GPCRs. A number of adhesion GPCR homologs identified in Fungi and Amoebozoa usually have very short N-terminus without the GAIN domain (Krishnan et al., 2012), which may be attributed to partial gene deletion. GAIN in PKD proteins was found in *Nematostella vectensis* (a sea anemone, e.g. Genbank: XP_001640030.1) and *Trichoplax adhaerens* (a placozoan, e.g. Genbank: XP_002110052.1). The GAIN domain is missing from the land plant lineage. Nup98 C-terminal domain was found in all major branches of eukaryotes. In contrast, FIIND domain and DUF1191 are limited to Chordata and Viridiplantae, respectively.

Discussion

In this work, I established remote homologous relationships between GAIN domain subdomain B, ZU5 domain and Nup98 C-terminal domain. By transitive PSI-BLAST homology searches, I discovered a diverse group of bacterial and archaeal domain sequences homologous to ZU5 domain, suggesting a common ancient origin of these domains. The ancestor domains may be extracellular bacterial homologs and retain the HFS motif. The cellular localization and autoproteolytic ability both evolved separately, resulting in distinct families that have functions specific to certain domain contexts and molecules.

The common autoproteolytic mechanism involves deprotonation and activation of the serine, threonine or cysteine by a general base (usually a histidine) which is followed by nucleophilic attack at the preceding peptide bond (Perler et al., 1997). The cleavage sites of GAIN and Nup98 are located at a sharply kinked loop between two strands in the opposite side of the beta-sandwich, which is stabilized by anchoring the hydrophobic side chain of the second residue in the motif (phenylalanine in Nup98 and leucine in BAI3) into a hydrophobic pocket. The scissile peptide bond either adopts a distorted *trans* conformation (Arac et al., 2012) or even a *cis* conformation (Sun & Guo, 2008), and such structural constraints facilitate N-O(S) acyl shift in autoproteolysis. In the evolution from the ancient ZU5 ancestor domain to current diverse branches of domain families, autoproteolysis also developed over time. Most human proteins containing the canonical ZU5 domain have lost the crucial S/T residue at the autoproteolytic site except for PIDD, UNC5C-like protein and MACC1, but the structural feature of the kink persists in available structures of UNC5B and ankyrins. Putative orthlogs of human UNC5C protein are found in Bilateria and Cnidaria (e.g. Genbank: XP_001638664.1 from *Nematostella vectensis*) by BLAST, while UNC5C-like orthlogs are only detected in Bilateria. The sequence in *Nematostella* likely resembles the common ancestor of UNC5 family as it preserves the HFS motif and also contains immunoglobulin-like and thrombospondin type 1 repeats domains in N-terminal extracellular region. Then during evolution, the extracellular domains were lost in UNC5C-like while the autoproteolytic sites were deteriorated in other human UNC5 members. In the case of UNC5 and ankyrin, ZU5 domain has gained other functions as a protein-protein interaction module utilizing different interfaces, such as autoinhibition of DD domain (R. Wang et al., 2009) and binding to spectrins (Ipsaro & Mondragon, 2010; C. Wang et al., 2012). In the ZU5 domain of the FIIND family, CARD8 and NLRP1 adapt an alternative method for autoproteolysis with a conserved site of SF↓S, and a nearby histidine likely participates in activation instead of the canonical histidine residue in the motif (D'Osualdo et al., 2011; Finger et al., 2012). However, the previous structural model of CARD8 was less reliable in the region around the proposed substituting histidine, and the side chain of that histidine 270 was shown pointing away from the cleavage site (D'Osualdo et al., 2011). With our multiple sequence alignment, the CARD8 ZU5-FIIND domain model constructed by MODELLER (Webb & Sali, 2014) places the histidine 270 side chain in proximity of the catalytic serine 297, supporting the hypothesis that both CARD8 and NLRP1 use an alternative histidine close in space to activate serine for autoproteolysis. These two histidines of CARD8 and NLRP1 (in beta-strand 8) are also aligned (Fig. 3.2, highlighted with a grey background). Interestingly, the ZU5 domain of MACC1 contains a DLS motif and a histidine at the equivalent positions in the sequence alignment (Fig. 3.2), which could potentially be an autoproteolytic site similar to those in CARD8 and NLRP1. As for Nup98, the structure has changed significantly while preserving the HFS motif. In the GAIN domain of adhesion GPCRs, the cleavage site is usually HLT/S. CILPs could have lost the autoproteolysis activity based on substitutions in the motif (Fig. 3.2). However, they are still regulated by proteolysis. It is possible that the predicted furin cleavage site (Lorenzo et al., 1998) is later inserted in CILP in place of the lost autoproteolytic capability.

Among all available structures, the ZU5 domain in ZO-1 is particularly intriguing because it terminates just before the scissile peptide bond (Fig. 3.2). When the peptide from its binding partner GRINL1A (glutamate receptor, ionotropic, N-methyl-D-aspartate-like 1A combined protein) is concatenated at the C-terminus, it forms a beta-strand that resembles the final strand after cleavage site in other available ZU5 structures and beta-strand 5 (magenta in Fig. 3.1(f)) interacts with it together with beta-strand 6 (green in Fig. 3.1(f)). This interaction is thought to be analogous to that between ZO-1 and MRCKβ (Huo et al., 2011). In the absence of a binding partner, beta-strand 5 folds down and pairs with beta-strand 6 occupying the interface in a closed conformation as shown in Fig. 3.1(e), or could form a flexible loop (Huo et al., 2011). One molecular basis for autoproteolytic functions could be to release this interaction site. The cleaved peptides all remain associated in GAIN, Nup98, PIDD, CARD8 and NLRP1 (Arac et al., 2012; D'Osualdo et al., 2011; Finger et al., 2012; Hodel et al., 2002; Tinel et al., 2007), which could compete with their binding partners. Indeed, Nup98 interacts with a loop in the beta-propeller domain in Nup82 by substituting and releasing the cleaved peptide (Stuwe, von Borzyskowski, Davenport, & Hoelz, 2012; Yoshida et al., 2011). The yeast Nup116, a paralog of Nup98, does not contain the autoproteolytic site, but its C-terminus terminates only four residues downstream of the equivalent cleavage position and also binds to Nup82 in the same way (Yoshida et al., 2011). Moreover, I discovered some homologous Nup98 sequences that end with the HF motif (e.g. Genbank: XP_002677377.1 from *Naegleria gruberi* and XP_002911255.1 from *Coprinopsis* *cinerea*) lacking the last beta-strand, just like ZO-1. Interestingly, one of the exon boundaries of human UNC5C-like protein falls exactly between the phenylalanine and serine in the autoproteolytic motif. Taken together, it demonstrates a possible evolutionary path of the ancient ZU5-like domain that intronization or loss of the exon after the cleavage site resulted in loss of the last beta-strand.

The functions of two autoproteolytic events in PIDD were shown to be crucial for regulating PIDD signaling in response to DNA damage (Tinel et al., 2007). Autoproteolysis at the first site removes the inhibitory N-terminal leucine rich repeats and is required for translocation to the nucleus. The resulting fragment PIDD-C can activate NF-KB, and the fragment PIDD-CC generated by the second autoproteolysis event can activate caspase-2 pathway alternatively. The function of the bipartite separation of N-terminal extracellular fragment and C-terminal fragment in adhesion GPCRs is still not clear (Langenhan et al., 2013). Several studies suggested that the cleaved N-terminal fragments in latrophilin-1 and EMR2 behave like independent proteins and reassociate with their C-terminal fragments upon ligand binding (Y. S. Huang et al., 2012; Volynski et al., 2004). Chimeric receptors with fragments from latrophilin-1, EMR2 and GPR56 were also observed by immunoprecipitation assays (Silva, Lelianova, Hopkins, Volynski, & Ushkaryov, 2009). But another study with latrophilin-1 in *Caenorhabditis elegans* proposed that only the structural integrity is required for normal signaling, but not the autoproteolysis of GAIN domain (Promel et al., 2012). Furthermore, in adhesion GPCR and PKD proteins, the autoproteolytic site has occasionally deteriorated, such as EMR1 and PKDL1 (Fig. 3.2). It is possible that the cleavage of the N-terminal fragment has family-specific functions or might even have evolved for reasons other than signaling, such as protection from mechanical stress in the SEA domain with autoproteolytic activity (Macao et al., 2006). In general, GAIN, ZU5 and Nup98-C domain could serve as a protein-protein interaction platform. In Nup98 and Nup116, the cleavage creates a binding site for interacting with other nucleoporins (Stuwe et al., 2012; Yoshida et al., 2011). Different adhesion GPCRs could exchange their N-terminal extracellular domains after autoproteolysis at the common GPS motif (Y. S. Huang et al., 2012; Volynski et al., 2004). ZO-1, lacking the last beta-strand, also uses the same region without the need of cleavage (Huo et al., 2011). For other domains that lost the cleavage motif like those in UNC5B and ankyrin, they evolved to utilize other interfaces for intermolecular and intramolecular domain interactions (Ipsaro & Mondragon, 2010; C. Wang et al., 2012; R. Wang et al., 2009).

Methods

Structure comparison were performed with Dali Server (Holm & Rosenstrom, 2010) using known structures of GAIN, ZU5 and Nucleoporin98 C-terminal domains in protein databank (PDB) (Berman et al., 2000). HHpred server was used for sensitive homology detection against the Pfam database and the PDB70 database (Soding et al., 2005). Transitive PSI-BLAST (Altschul et al., 1997) searches were conducted to find sequence homologs. PSI-BLAST hits were first clustered by BLASTCLUST with score coverage threshold 0.5 (-S option). Representative sequences of each cluster were used to initiate new PSI-BLAST searches. Such a procedure was repeated until convergence. The transitive PSI-BLAST search was performed for GAIN, ZU5, Nup98-C sequences derived from structures and remote homologs detected by HHpred.

Human proteins found in transitive PSI-BLAST searches were selected including several GPCR sequences. Multiple sequence alignment was built by PROMALS3D (Pei, Kim, & Grishin, 2008) and edited manually. The select proteins were also subjected to domain architecture analysis. CD-Search (Marchler-Bauer et al., 2011), HMMERSCAN (Finn, Clements, & Eddy, 2011) and HHpred (Soding et al., 2005) were used to detect conserved domains in CDD (Marchler-Bauer et al., 2011) and Pfam (Finn et al., 2014) databases with default parameters. Signal peptides and transmembrane helices were predicted by SignalP (Petersen, Brunak, von Heijne, & Nielsen, 2011) and Phobius (Kall, Krogh, & Sonnhammer, 2007).

All homologous sequences were collected and clustered by CD-HIT (Fu, Niu, Zhu, Wu, & Li, 2012) at 80% sequence identity. The reduced set was clustered and visualized graphically by CLANS (Frickey & Lupas, 2004) in two-dimensional space where all-against-all BLAST were run and the negative logarithm of pairwise p-values (small than 1e-4) were used as attractive forces to reach a layout in equilibrium.

Homology model of CARD8 ZU5 domain was built by MODELLER (Webb & Sali, 2014) using default parameters. The multiple sequence alignment of ZU5 domains from CARD8, UNC5B (PDB: 3G5B), ankyrin (PDB: 3UD1) and GAIN domain from BAI3 (PDB: 4DLO) was given as the input.

Figures

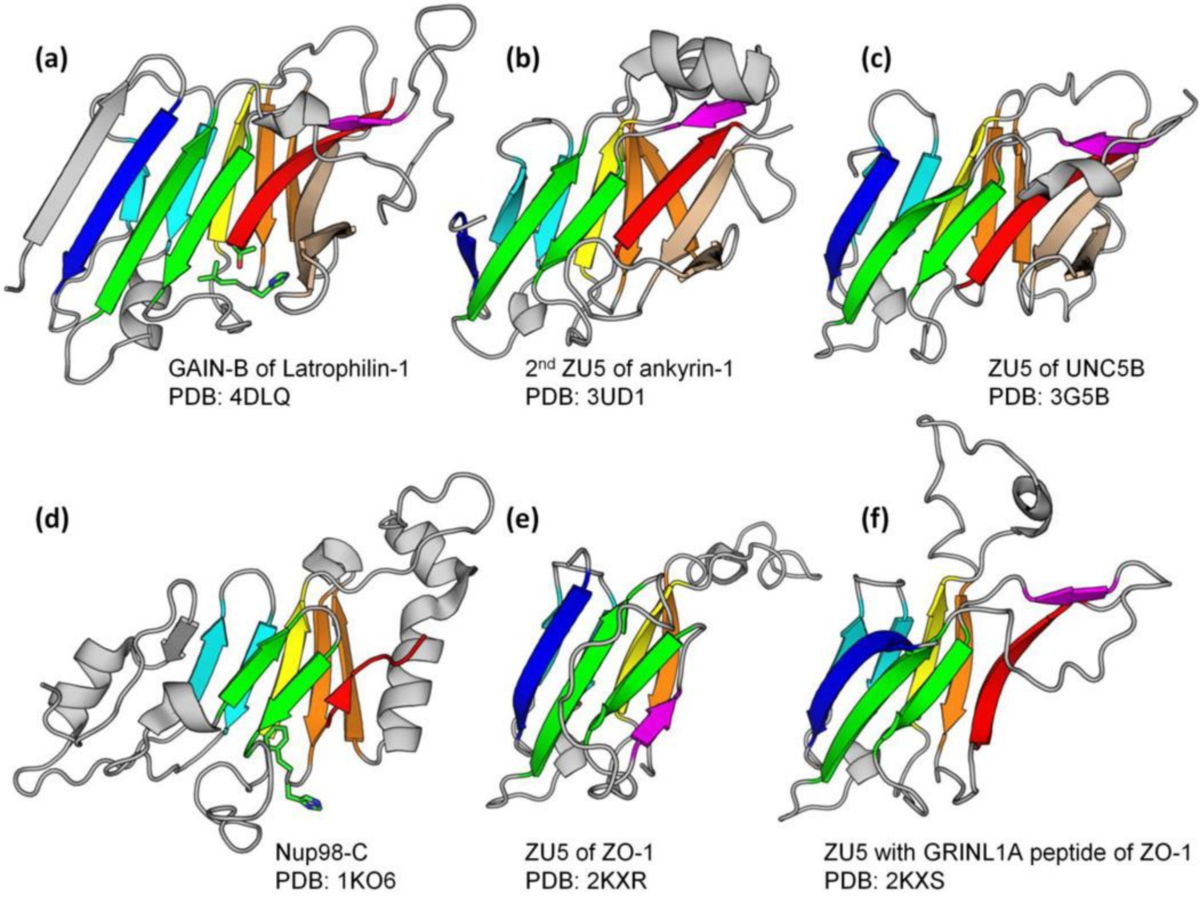
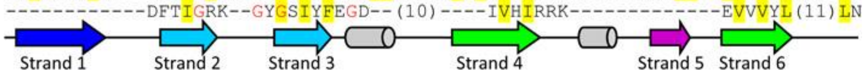


Fig. 3.1. Structures of GAIN-B, ZU5 and Nup98 C-terminal domain.

GAIN domain subdomain B in latrophilin-1, ZU5 domains in ankyrin-1, UNC5B and ZO-1 and Nucleoporin98 C-terminal domain are superimposed and rendered in cartoon by Pymol. Core strands are colored generally from blue to red with paired strands in a hairpin colored identically. The side chains of autoproteolytic sites in Nup98 and GAIN are shown in sticks. Note that the serine in Nup98 structure is missing.

```
NP_001695.1|BAI3*           702 VVASIQKLPA(6)INFPMK(15)DRVVIPKSIF(12)VFVLGAVLYKNL(7)RNYTVINSKIIVVTI(6)-TD
NP_001008701.1|LPHN1*       679 VVLEVTVLNT(6)LVFPQE(4)KNSIQLSAKTI(6)GVVKVVFILYNN(24)GASLVVNSQVIAASI(8)-LM
NP_001965.3|EMR1            433 LDIESKVINK(6)VTLDLVA-KGDKMKIGCSTI(6)ETTGVAFVSFVG(20)EIKLKMNSRVVGGIM(7)-FS
NP_001009944.2|PKD1        2781 IVAQGKRSDPRSLLCYGGAPGPGCHFSIPEAFS(7)DVVQLIFLVDSNP(5)ISNYTVSTKVASMAF(14)SE
NP_612152.1|PKD1L1         1467 MEFRTLLHYNL-QSSVQSL--GSVQVHLPGDLA(12)CYISQLILFKKN(4)SQAPGQIGGVVGLNL(13)LR
NP_001265354.1|PKD1L2       382 ISVYTNRIQP(5)SSLRPDAADSATFMLPAASS(8)EPVDIKIMSFPKS(4)RSHFDVSGTVGGLRV(12)LS
NP_853514.1|PKD1L3          418 ATLLLSRQNI(5)SSYTLGHPAPVRLGFPSALA(8)PGVNVQITGLAFN(4)LDNRNIVGSIGSVLL(11)LM
NP_006062.1|PKDREJ          859 FNMYVKKVEKWGINQLFRNEKHCRNCFYPTLNV(9)PPISTMFCDFTND(5)NDQENTSVEVSGFRM(12)TP
NP_003248.3|ZO-1*          1633 VATARGIFNSN-GGVLSSIE-TGVSIIIPQGAIPEGVEQEIYFKVCRDN(8)EKGETLLSPLVMCGPHGLKFL
NP_588610.2|UNC5A           440 SNMTYGTFNFL-GGRLMIPN-TGISLLIPPDAIPRGKIYEIYLTLHKPE(6)AGCQTLLSPIVSCGPPGVLLT
NP_734465.2|UNC5B*          542 GSSVSGTFGCL-GGRLSIPG-TGVSLLVPNGAIPQGKFYEMYLLINKAE(6)EGTQTVLSPSVTCGPTGLLLC
NP_775832.2|UNC5CL          101 LVFSAREVDHR-GGCLMLQD-TGISLLIPPGAVAVGRQERVSLILVWDL(6)SQAQGLVSPVVACGPHGASFL
NP_065209.2|ANK1*           912 GFLVSFMVDAR-GGSMRGSRHNGLRVVIPPRTCA--APTRITCRLVKPQ(8)AEEEGLASRIIALGPTGAQFL
NP_065209.2|ANK1*          1070 -CQDYDTIGPE-GGSLKSKLVPLVQATFPENAVT--KRVKLALQAQPVP(7)LGNQATFSPIVTVEPRRRKFH
NP_001164171.1|DTHD1        166 NIMEKEYLDV--LSDVTGPQ-VSCYITAPSYVL(6)IINHMSSLIVGD------NEELVSNVITIEC(6)-VP
NP_001164171.1|DTHD1        302 -KKESFTVTKK-GLALKSSMDSRISLNYPPGVFT--SPVLVQLKIQPV(14)FYSVQSTSPLIHIQH(5)-FQ
NP_055336.1|SH3BP4          316 ETNIVCKLDSS-GGAVQLPD-TSISIHVPEGHVAPGETQQISMKALLDP(5)SDRSCSISPVLEVKLSNLEVK
NP_877439.3|MACC1           211 EVTIACKVNHQ-GGSVQLPE-SDITVHVPQGHVAVGEFQEVSLRAFLDP(5)HDLSCTVSPLLEIMLGNLNTM
NP_665893.2|PIDD            321 SDLDSFPVTPQ-GCSVTLA--CGVRLQFPAGATA--TPITIRYRLLLPE(6)GPHDALLSHVLELQPHGVAFQ
NP_665893.2|PIDD            455 -VSNACLVPPE-GTLLCSSGHPGVKVIFPPGATE--EPRRVSMQVVRMA(8)GEPEAAVSPLLCLSQ(5)-FL
NP_001171829.1|CARD8        177 TNRYSVWFPT--AGWYLWSA-TGLGFLVRDEVT-----VTIAFGSWSQH(7)HEQWLVGGPLFDVTAEP--EE
NP_127497.1|NLRP1          1095 KNLYRVHFPV--AGSYRWPN-TGLCFVMREAVT-----VEIEFCVWDQF(6)QHSWMVAGPLLDIKAEP---G
NP_003604.3|CILP-1          579 LEAMETNIIP--LGEVVGED-PMAELEIPSRSF(9)GKVKASVTFLDP(20)DTFPLRTYGMFSVDF(9)-NA
NP_001010924.1|FAM171A1     125 LMVYEDVVQI--VSGFQGAR-PQPRVHFQRRAL(8)SDLTAFLTAASS(21)TRHDLTPVTAVSVHL(9)-VD
NP_005378.4|NUP98*          764 ------------DFTIGRK--GYGSIYFEGD--(10)----IVHIRRK------------EVVVYL(11)LN
```



```
NP_001695.1|BAI3*           SFLEIELAHLANGTLNPYCVLWDD(6)LGTWSTQG---CKTVLTAD--SHTKCLCDRLSTFAILAQQ   865[1522]
NP_001008701.1|LPHN1*       DPVIFTVAHL(4)HFNANCSFWNY(5)LGYWSTQG---CRLESKN---THTTCACSHLTNFAVLMAH   847[1474]
NP_001965.3|EMR1            DPIIYTLENI(5)FERPICVSWSTDVKGGRWTSFG---CVIEAES---TYTICSCNQMANLAVIMAS   593[ 886]
NP_001009944.2|PKD1         RAITVKVPNN(132)TSLCQYFSEE--DMVWRTEG---LLPLEEPTS-RQAVCLTRHLTAFGASLFV  3057[4303]
NP_612152.1|PKD1L1          KPVMVEFGEE(123)WIRCLFWDK----REWKSER---FSPQPGTSP-EKVNCSYHRLAAFALLRRK  1731[2849]
NP_001265354.1|PKD1L2       ENIEILLPRH(113)LSHCVFWDEV--QETWDDSG---CQVGPRPTS-YQTHCLCNHLTFFGSTFLV   639[1774]
NP_853514.1|PKD1L3          EDIEIMLWRN(94)AVTQCYYWEIH--NQTWSSAG---CQVGPQIST-LRTQCLCNHLTFFASDFFV   677[1732]
NP_006062.1|PKDREJ          DVAEVYLVRK(143)SVQCLDMYGI--QSEWREGY---CILGEKTSW-YEVHCICKHTHYVMAKVIV  1165[2253]
NP_003248.3|ZO-1*           KPVELRLPHC(20)-------------------------------------NCVSVLIDHF---------  1748[1748]
NP_588610.2|UNC5A           RPVILAMDHC(4)PDSWSLRLKKQSC-EGSWEDV(11)YYCQLEA---SACYVFTEQLGRFALVGEA   584[ 842]
NP_734465.2|UNC5B*          RPVILTMPHC(4)ARDWIFQLKTQAH-QGHWEEV(11)CYCQLEP---RACHILLDQLGTYVFTGES   686[ 945]
NP_775832.2|UNC5CL          KPCTLTFKHC(4)SHARTYSSNTTLLDAKVWRPLG--RPGAHASR---DECRIHLSHFSLYTCVLEA   237[ 518]
NP_065209.2|ANK1*           SPVIVEIPHF(5)GDRELVVLRSEN--GSVWKEHR(28)---------RVCRIITTDFPYFVIMSRL  1069[1881]
NP_065209.2|ANK1*           RPIGLRIPLP(15)TSLRLLCSVI(4)QAQWEDIT-GTTKLVYAN---ECANFTTNVSARFWLSDCP  1217[1881]
NP_001164171.1|DTHD1        FPIGIAIPFT(5)NYRDIMVKVCDINLQSSYLNPN(5)MKGGYKG---TCASVKVYKLGIFSVVSCL   301[ 781]
NP_001164171.1|DTHD1        KPVTLFLPCS(46)ECLKLLGFRSQ--DSGWCGLDD--VVKTIQS---GLVSVELYHLERFIVLHLS   484[ 781]
NP_055336.1|SH3BP4          TSIILEMKVS(12)VGLQCLRSDSK--EGPYVSV----PLNCSCG---DTVQAQLHNLEPCMYVAVV   454[ 963]
NP_877439.3|MACC1           EALLLEMKIG(12)TEMVCLHSLGK--EGPFKVL----SNCYIYK---DTIQVKLIDLSQVMYLVVA   349[ 852]
NP_665893.2|PIDD            QDVGLWLLFT(4)RRCREVVVRTRN--DNSWGDLE-TYLEEEAPRQ--LWAHCQVPHFSWFLVVSRP   454[ 910]
NP_665893.2|PIDD            QPVTVQLPLP(10)SRLHLLYWAPP--AATWDDIT-AQVVLELTH---LYARFQVTHFSWYWLWYTT   596[ 910]
NP_001171829.1|CARD8        AVAEIHLPHF(10)SWFLVAHFKN----EGMVLE----HPARVEP---FYAVLESPSFSLMGILLRI   305[ 537]
NP_127497.1|NLRP1           AVEAVHLPHF(10)SLFQMAHFKE----EGMLLE----KPARVEL---HHIVLENPSFSPLGVLLKM  1221[1473]
NP_003604.3|CILP-1          GKVKVHLDST(7)ISTVKLWSLNPD--TGLWEEEGD--FKFENQR(9)LVGNLEIRERRLFNLDVPE   748[1184]
NP_001010924.1|FAM171A1     GPIYVTVPLA(7)NAYVAAWRFDQK--LGTWLKSGL--GLVHQEQGS-LTWTYIAPQLGYWVAAMSP   287[ 890]
NP_005378.4|NUP98*          RKAEVTLDGV(34)------------------------AQFEYRTPE-GSWVFKVSHFSKYGLQDSD   889[ 937]
```

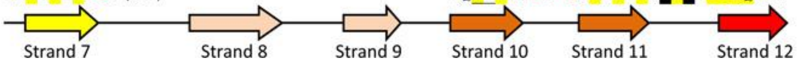


Fig. 3.2. Multiple sequence alignment of selected human proteins containing GAIN, ZU5 and Nup98 C-terminal domains.

Human proteins found in transitive PSI-BLAST searches were selected including several GPCR sequences. Multiple sequence alignment was built by PROMALS3D (Pei & Grishin, 2014) and edited manually. Common gene names following NCBI Genbank accession numbers are used for each protein. Names of GAIN, canonical ZU5, ZU5-FIIND, ZU5-CF and Nup98 C-terminal domain are in red, black, blue, magenta and cyan colors, respectively. An asterisk is labeled where the protein or one of its close homologs has available structures. The conserved serine, threonine and histidine in the cleavage site are highlighted in black background and putative active site histidines in beta-strand 8 in ZU5-FIIND and MACC1 are shaded in grey. The protein length is noted in brackets at the end of the alignment. Nonpolar residues in mainly hydrophobic positions are highlighted in yellow. Glycines and prolines are colored in red. The column of mainly aromatic residue is in bold. Secondary structures are represented as arrows (beta-strands) and tubes (alpha-helices) below and colored consistently with structures in Fig. 3.1. Insertions are represented by the numbers of inserted residues in parentheses. Several one residue insertions are black underscored and the red underscore in PKDREJ represents a 19-residue insertion before the deteriorated cleavage motif. And the insertion in CILP-1 (sequence: RNKREDRTF) with a consensus furin-like cleavage site is colored in blue.

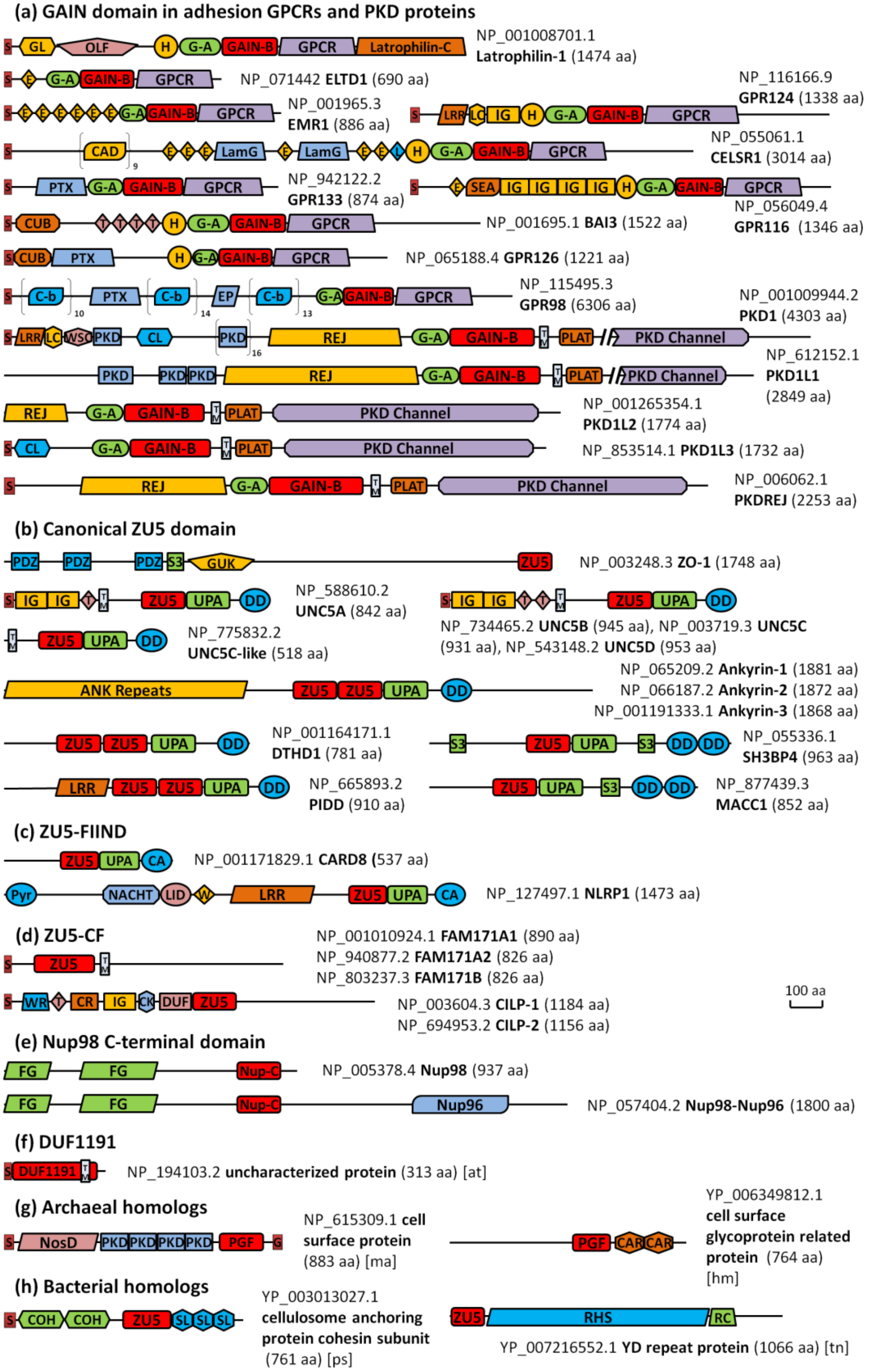

(a) GAIN domain in adhesion GPCRs and PKD proteins
S GL OLF H G-A GAIN-B GPCR Latrophilin-C
NP_001008701.1 Latrophilin-1 (1474 aa)
S E G-A GAIN-B GPCR
NP_071442 ELTD1 (690 aa)
NP_116166.9 GPR124 (1338 aa)
S E E E E E E E G-A GAIN-B GPCR
NP_001965.3 EMR1 (886 aa)
S LRR LC IG H G-A GAIN-B GPCR
S CAD 9 E E E LamG E LamG E E L H G-A GAIN-B GPCR
NP_055061.1 CELSR1 (3014 aa)
S PTX G-A GAIN-B GPCR
NP_942122.2 GPR133 (874 aa)
S E SEA IG IG IG IG H G-A GAIN-B GPCR
NP_056049.4 GPR116 (1346 aa)
S CUB T T T T H G-A GAIN-B GPCR
NP_001695.1 BAI3 (1522 aa)
S CUB PTX H G-A GAIN-B GPCR
NP_065188.4 GPR126 (1221 aa)
S C-b 10 PTX C-b 14 EP C-b 13 G-A GAIN-B GPCR
NP_115495.3 GPR98 (6306 aa)
NP_001009944.2 PKD1 (4303 aa)
S LRR LC WSC PKD CL PKD 16 REJ G-A GAIN-B TM PLAT PKD Channel
PKD PKD PKD REJ G-A GAIN-B TM PLAT PKD Channel
NP_612152.1 PKD1L1 (2849 aa)
REJ G-A GAIN-B TM PLAT PKD Channel
NP_001265354.1 PKD1L2 (1774 aa)
S CL G-A GAIN-B TM PLAT PKD Channel
NP_853514.1 PKD1L3 (1732 aa)
S REJ G-A GAIN-B TM PLAT PKD Channel
NP_006062.1 PKDREJ (2253 aa)
(b) Canonical ZU5 domain
PDZ PDZ PDZ S3 GUK ZU5
NP_003248.3 ZO-1 (1748 aa)
S IG IG T TM ZU5 UPA DD
NP_588610.2 UNC5A (842 aa)
S IG IG T T TM ZU5 UPA DD
NP_734465.2 UNC5B (945 aa), NP_003719.3 UNC5C (931 aa), NP_543148.2 UNC5D (953 aa)
TM ZU5 UPA DD
NP_775832.2 UNC5C-like (518 aa)
ANK Repeats ZU5 ZU5 UPA DD
NP_065209.2 Ankyrin-1 (1881 aa)
NP_066187.2 Ankyrin-2 (1872 aa)
NP_001191333.1 Ankyrin-3 (1868 aa)
ZU5 ZU5 UPA DD
NP_001164171.1 DTHD1 (781 aa)
S3 ZU5 UPA S3 DD DD
NP_055336.1 SH3BP4 (963 aa)
LRR ZU5 ZU5 UPA DD
NP_665893.2 PIDD (910 aa)
ZU5 UPA S3 DD DD
NP_877439.3 MACC1 (852 aa)
(c) ZU5-FIIND
ZU5 UPA CA
NP_001171829.1 CARD8 (537 aa)
Pyr NACHT LID W LRR ZU5 UPA CA
NP_127497.1 NLRP1 (1473 aa)
(d) ZU5-CF
S ZU5 TM
NP_001010924.1 FAM171A1 (890 aa)
NP_940877.2 FAM171A2 (826 aa)
NP_803237.3 FAM171B (826 aa)
S WR T CR IG CK DUF ZU5
NP_003604.3 CILP-1 (1184 aa)
NP_694953.2 CILP-2 (1156 aa)
100 aa
(e) Nup98 C-terminal domain
FG FG Nup-C
NP_005378.4 Nup98 (937 aa)
FG FG Nup-C Nup96
NP_057404.2 Nup98-Nup96 (1800 aa)
(f) DUF1191
S DUF1191 TM
NP_194103.2 uncharacterized protein (313 aa) [at]
(g) Archaeal homologs
S NosD PKD PKD PKD PKD PGF G
NP_615309.1 cell surface protein (883 aa) [ma]
PGF CAR CAR
YP_006349812.1 cell surface glycoprotein related protein (764 aa) [hm]
(h) Bacterial homologs
S COH COH ZU5 SL SL SL
YP_003013027.1 cellulosome anchoring protein cohesin subunit (761 aa) [ps]
ZU5 RHS RC
YP_007216552.1 YD repeat protein (1066 aa) [tn]

Fig. 3.3. Domain architecture diagrams of selected proteins containing GAIN, ZU5, Nup98 C-terminal domains.

The domain architecture diagrams are drawn roughly to scale, except that two long PKD family proteins have part of the PKD channel domain cut out which is indicated by double slashes and the consecutive domain repeats in CELSR1, GPR98 and PKD1 are shown in parentheses with repeat number labeled. Representative adhesion GPCRs are selected, and others were summarized in review (Lagerstrom & Schioth, 2008). CD-Search (Marchler-Bauer et al., 2011), HMMERSCAN (Finn et al., 2011) and HHpred (Soding, 2005) were used to detect conserved domains in CDD and Pfam databases with default parameters. Signal peptides, transmembrane helices and GPI anchors were predicted by SignalP (Petersen et al., 2011), Phobius (Kall et al., 2007) and PredGPI (Pierleoni, Martelli, & Casadio, 2008). NCBI Genbank accession numbers and protein lengths are annotated for each protein. Domain name abbreviations are listed as follows: ANK: ankyrin repeats; CA: caspase recruitment domain, CARD; CAD: Cadherin domain; CAR: CARDB, cell adhesion related domain found in bacteria; CK: cystine knot-like domain; CL: C-type lectin domain; COH: cohesin domain; CR: carboxypeptidase regulatory-like domain; CUB: CUB (for complement C1r/C1s, Uegf, Bmp1) domain; C-b: Calx-beta domain; DD: death domain; DUF: DUF4480; E: Calcium-binding EGF domain; EP: Epitempin/Epilepsy-associated repeats; FG: FG and GLFG motif repeat region. Nup98 has totally 39 FG repeats including 9 GLFG motifs; G: GPI anchor; G-A: GAIN subdomain A; GAIN-B: GAIN subdomain B; GL: galactose binding lectin domain; GUK: guanylate kinase homologues; H, hormone receptor domain; IG: Immunoglobulin domain; L: Laminin EGF-like domain; LamG: Laminin G domain; Latrophilin-C: latrophilin cytoplasmic C-terminal region; LC: leucine rich repeat C-terminal domain; LID: AAA+ ATPase lid domain; LRR, leucine rich repeat; NosD: periplasmic copper-binding protein (NosD); Nup-C: nucleoporin98 C-terminal domain; Nup96: nuclear protein 96; OLF: olfactomedin-like domain; PGF: PGF_pre_PGF domain; PKD: polycystic kidney disease I (PKD) domain; PLAT: PLAT (Polycystin-1, Lipoxygenase, Alpha-Toxin) or LH2 (Lipoxygenase homology 2) domain; PTX: pentraxin domain; Pyr: pyrin domain; RC: RHS repeat-associated core domain; REJ: receptor for egg jelly domain; RHS: RHS (rearrangement hotspot) or YD repeats; S, signal peptide; SEA: SEA (found in Sea urchin sperm protein, Enterokinase, Agrin) domain; SL: S-layer homology (SLH) domain; S3: SRC homology 3 domain; T: thrombospondin type 1 repeats; TM: transmembrane helix; UPA: UPA (common in UNC5, PIDD, Ankyrin) domain; W: winged helix-turn-helix domain; WSC: WSC domain and WR: mucin-2 protein WxxW repeating region. Organisms other than Homo sapiens are labeled in square brackets with the following abbreviations: at: Arabidopsis thaliana; hm: Haloferax mediterranei ATCC 33500; ma: Methanosarcina acetivorans C2A; ps: Paenibacillus sp. JDR-2; tn: Thioalkalivibrio nitratireducens DSM 14787.

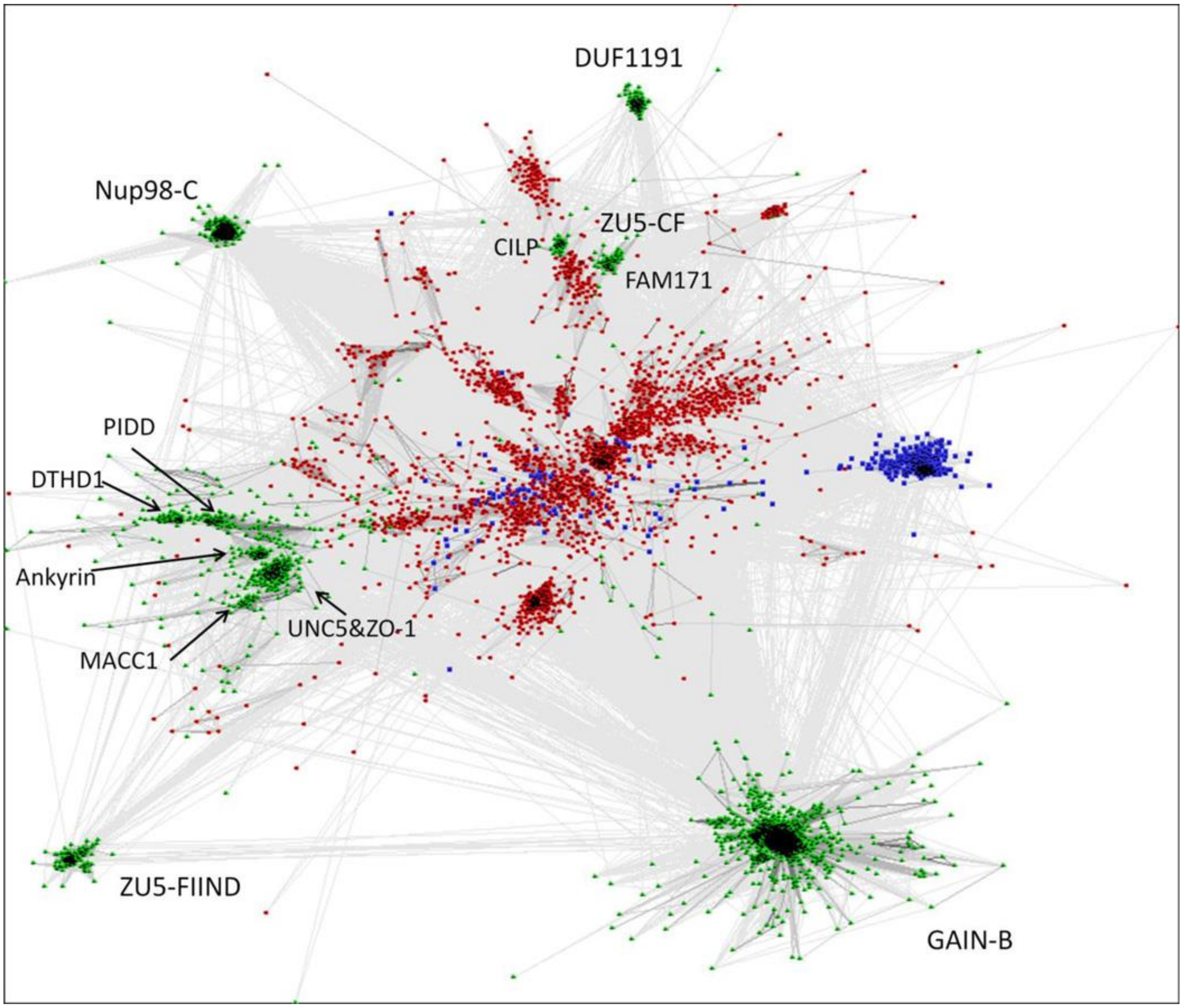


Fig. 3.4. Sequence clustering of all GAIN, ZU5 and Nup98-C homologs.

Nonredundant homologous domain sequences were first clustered by CD-HIT (Fu et al., 2012) at 80% sequence identity. Representatives were then clustered and visualized by CLANS (Frickey & Lupas, 2004) in two-dimensional space. Sequences are colored by domains of life, with green used for Eukaryote, red for Bacteria and blue for Archaea. Protein family names are labeled. Connections are drawn for links with BLAST p-value less than 1e-4.

CHAPTER FOUR

Sequence families built on ECOD structural domains

Introduction

ECOD classifies protein domains based on their evolutionary history and groups remote homologs that share common ancestors in the same Homology group (H-group) while recognizing fine clustering of close homologs by families (F-group) (Cheng et al., 2014). Distant homologs that have diverged significantly in evolution may have altered sequence so much that it is beyond the sensitivity of current sequence-based homology detection programs, or they could even have evolved with different topologies, which is characterized by ECOD Topology group. Sequence families were introduced to represent a group of proteins that are highly similar to each other and usually contain some conserved residues and motifs with implication of function or structural interaction. The sequences in a family are usually aligned, and a hidden Markov model (HMM) is derived from the multiple sequence alignment to represent the family for search and domain annotation.

Multiple concepts comprising domain definitions, *i.e.* functional, structural, and homology-based, lead to different perspectives between different types of protein classifications. Whereas structural classifications may have clearer domain boundaries, sequence-based classifications can access larger datasets that more comprehensively sample protein space (including entire genomes). In ECOD, I consider domains as independent evolutionary units and manually curate the domain boundary for structural representatives. Families in ECOD were primarily dependent on Pfam database (Cheng et al., 2014;

Schaeffer, Liao, Cheng, & Grishin, 2017). Although Pfam recently expedited their production (Finn et al., 2016), not all structures in the PDB database can be found in Pfam, especially the ones deposited lately. Another intrinsic difference between existing sequence family databases and ECOD is that ECOD domain boundaries also take account into structural information. This naturally leads to disagreement between domains defined in ECOD and those defined by sequence databases (Schaeffer, Kinch, Liao, & Grishin, 2016). One ECOD domain could be covered by several Pfam domains or the other way around, which poses a challenge for consistent classification in ECOD.

Firstly, I mapped Pfam version 28 families to ECOD 40% sequence redundant representative set and compared domain boundaries to demonstrate the degree of the problem. 25,989 (81.7%) domains are mapped to exactly one Pfam family; 1,334 (4.2%) domains contain multiple non-overlapping mappings to distinct Pfam families, indicating potential Pfam families to merge; 4,476 (14.1%) domains have no significant Pfam hits. For the one-to-one mapping, the coverage of Pfam families by ECOD domain exhibits an exponential distribution and 91.2% of these domains have more than 50% coverage by a Pfam family (Fig. 4.1(a)). Those ECOD domains that can be classified by Pfam are consistent with the Pfam domain definition. ECOD domains with low coverage by Pfam families can be attributed to continual internal repeats (e.g., β-propeller, ARM repeat, β-helix, etc.), where Pfam only defines one or several repeating units as a family and ECOD tends to cover all. Others are usually explained by the nature of slower evolution of structure (Murzin, 1998). The sequence family may only capture the most conserved core of the actual domain especially when it is established before any structural information is available, while individual structures can diverge in sequence space and develop assorted insertions and decorations. The percentage of Pfam overlap on ECOD domain suggests the degree to which structures diverge while keeping detectable conserved sequence signal. However, there are also some Pfam families that are very short and are better described as a conserved motif.

Out of 1,334 ECOD domains with non-overlapping regions that map to distinct Pfam families, 426 are mapped to unique Pfam domain arrangements where the individual families contain no structurally compact domain and do not occur independently. Often the co-occurrence of these families is noted by Pfam. For example, XPG_N (PF00752) and XPG_I (PF00867) were first discovered as two highly conserved N-terminal and internal regions of Xeroderma Pigmentosum Complementation Group G (XPG) proteins (Scherly et al., 1993). The XPG family includes various structure-specific nucleases, such as XPG/RAD2, flap endonuclease 1 (FEN1), and exonuclease 1 (EXO1) (Harrington & Lieber, 1994). Initially identified via comparison to yeast RAD2, XPG_N and XPG_I are separated by more than 600 amino acids in the alignment, but in FEN1 and EXO1 the spacer is shorter than 50 amino acids (Harrington & Lieber, 1994; Scherly et al., 1993). Later, crystal structures showed that XPG_N and XPG_I intertwine to form a compact α/β three-layered sandwich (Fig. 4.1(b)) (Hwang, Baek, Kim, & Cho, 1998; Mietus et al., 2014; Tsutakawa et al., 2011), which suggests that both belong to the large HAD-like superfamily (Burroughs et al., 2006). A number of protein domains in the XPG_I family incorrectly include the C-terminal SAM-like domain H2TH motif (Fig. 4.1(b,c)). The XPG family active site is located above the β-sheet where the N and I regions meet and is composed of carboxylate groups from both segments (Fig. 4.1(b)) (Tomlinson, Atack, Chapados, Tainer, & Grasby, 2010). These two families also co-occur with high frequency in Pfam. Therefore, I determine that XPG-N and XPG-I are best represented as a single family (Fig. 4.1(c), pink), where the H2TH motif belongs to a separate C-terminal domain (Fig.4.1(c), cyan). Conversely, ECOD domains sometimes split Pfam defined domains. Often, functional sites form at the intersection of structural domains. For such cases, sequence-based classifications tend to merge the structural domains into a single sequence domain due to similar conservation patterns that define the functional site. I found 771 Pfam families mapped to multiple ECOD domains in different H-groups in the 40% redundant set. The most commonly split H-groups are HTH, Rossmann-related, P-loop domain-related, and immunoglobulin-related, which are also the most populated groups generally in ECOD (Fig. 4.1(d)).

The definition of domain could evolve as our understanding of proteins advances. The inconsistency between Pfam family and ECOD domain reflects a need to improve current definitions of protein families that are purely based on sequence information by incorporating evolutionary insights from protein structures. Here I aim to build alignment and family profile from our ECOD domains, not only to help provide consistent family grouping within ECOD, but also to improve boundary definition of existing families with structural information.

Construction of family alignment and profile

The whole process is summarized as a flowchart shown in Fig. 4.2. The classification of ECOD sequence families and temporary solutions for domains that cannot be mapped to existing families were described in initial ECOD publication and recent updates (Cheng et al., 2014; Schaeffer et al., 2017). Briefly, I assigned Pfam version 27 families to classified ECOD domain when possible with HMMER 3.1b2 (Eddy, 2011). Unmapped domains were clustered and served as provisional sequence families.

Then for each ECOD F-group, which is either a provisional family or can be mapped to one or more non-overlapping Pfam families, I clustered domains to 70% sequence redundancy and ran all-to-all pairwise structure alignments for representatives with Dali (Holm & Park, 2000), TM-align (Zhang & Skolnick, 2005), and FAST (Zhu & Weng, 2005). The compiled list of the structure alignments was used as custom constraint for PROMALS3D (Pei & Grishin, 2014) to build multiple sequence alignment for all and non-redundant domains.

Next, I tried to define the boundary of the core of the alignment before building profiles. For larger groups, the start and end could be decided based on consensus gapness of the alignment. For small groups or even singletons, I resorted to a software SCR to predict structure core, which utilizes both sequence and structure information including secondary structure, contacts, sequence conservation, and etc. (I. K. Huang, Pei, & Grishin, 2013). This process also helps to remove non-homologous regions at the ends of domain such as linkers, expression tags, which could introduce contamination in profile construction. The performance of the core definition was evaluated by plotting the distribution of the percentage of the alignment that is cut. For most groups, the trimmed proportion is less than 20%, and if the percentage exceeds 50%, which are mostly derived from prediction results, I just decided to keep the alignment intact.

The trimmed core alignment was then converted to a profile and then the profile was searched against 80% redundancy Uniprot reference proteomes (The UniProt, 2017) with

HMMER to include sequences without structures, from which the seed HMM profile was built. A full alignment was also produced by searching the seed profile against the reference proteome database. This approach and the underlying sequence database are similar to what Pfam recently migrated to since version 28 (Finn et al., 2016).

I used HHalign in HHsuite (Soding, 2005) to pairwisely compare and score all family profiles in the same H-group and merged redundant families that are highly similar to each other. The scores were also converted to distances and then used to build phylogenic trees to show the relationship of homologous families on the website with the help of pHMM-Tree software (Huo et al., 2017).

Validation of alignment quality

Traditionally, evaluation of the quality of multiple sequence alignments uses established benchmarks of manual alignments or *ad hoc* structural alignments as gold standard (Pei et al., 2008; Thompson, Koehl, Ripp, & Poch, 2005; Van Walle, Lasters, & Wyns, 2005). I sought to evaluate our alignments generated by PROMALS3D with LGA program (Zemla, 2003). LGA is a program frequently used in model evaluation in CASP competition for the global distance test (GDT) and the total score (GDT_TS), which ranges from 0 to 100 and describes the average percentage of residues that can match under different distance thresholds. LGA can also run in sequence independent analysis mode if the sequence equivalence is given as input.

I utilized LGA to superimpose and score a pair of ECOD domains with the sequence equivalence defined in the sequence alignment and calculated the average GDT_TS score for each family. The distribution of the average GDT_TS score per family is compared between

ECOD alignments and Pfam alignments (Fig. 4.2(a)). In general, the two distributions are similar with peaks around GDT_TS score of 80. It suggests that the average quality of ECOD alignment which is built automatically with structural constraints is comparable to that of Pfam alignments which involve manual curation. A similar result was obtained with comparison of only those ECOD families containing PDBs in the Pfam dataset.

On the other hand, the distribution of Pfam alignments seems to have a longer tail on the left side with lower scores. In some hard cases, divergent family members could have various insertions and decorations at different locations, making alignment with only sequence information difficult. Such an example is shown in Fig. 4.2(b,c), where corresponding residues in the alignment are mapped on the structures with the same color. Pfam alignments are mostly continuous with few gaps in the middle, which actually represents registration shift in the alignment and thus has a very low GDT_TS score of 28.8 (Fig. 4.2(b)). The alignment built with Promols3D makes more gaps to take care of corresponding secondary structure elements and loops of differing lengths (Fig. 4.2(c)), which results in a much better GDT_TS score of 71.3.

ECOD family alignments should be of high quality on average based on structural evaluation criteria. In most cases, close homologs in a family tend to have the similar overall fold, except for those that have large flexible regions or can undergo large conformational changes, such as the N-terminal domain of chaperone SurA (PDB 3RFW and 3NRK).

Statistics and new families

In total, I have 11,653 families, each of which is defined by an alignment of sequences with structures, alignments of sequences from Uniprot reference proteome and a

HMMER profile. Firstly, I looked at the distribution of the number of sequences in ECOD families. The histogram of the natural logarithm of seed alignment size is plotted in Fig. 4.3(a), which shows a striking peak of small families. More specifically, the proportion of singleton family is 8.9%, and families with no more than 10 members occupy 26.5%. I calculated the same statistics for Pfam (Fig. 4.3(b)); the distribution exhibits a single peak, and the percentages of singleton family and families with no more than 10 members are 1.9% and 14.5%, respectively.

In order to explore what constitute the small families, I used HHsearch (Soding, 2005) to compare ECOD family profiles with the latest Pfam family profiles (Finn et al., 2016) and as well as CDD database (Marchler-Bauer et al., 2015), which incorporates many databases including Pfam and some curated families at NCBI. Depending on whether there are hits passing the threshold and the length of the hits, ECOD family is divided into four categories: family with no hits is referred to as “New family”; family with one hit of comparable length is referred to as ‘Identical family’; family with one hit of significantly different length is referred to as ‘modified family’; family with multiple non-overlapping hits is referred to as ‘merged family’. When I checked the sizes of new family comparing with Pfam, it shows a strong bias towards small families (Fig. 4.3(c)). Among these new families, 22.2% of them are singletons and 53.0% have no more than 10 members. Similar results were obtained when compared with CDD database, while less families are labeled new, but the percentage of small families are higher, with 32.8% for singletons and 71.1% for families no larger than 10. It suggests that the overrepresentation of small families is due to the sampling bias of structures deposited in PDB database, and they are mostly contributed by new families initiated by structures that cannot be found in existing families.

The numbers of the four categories of families when compared with Pfam are shown in Fig. 4.3(d). 36% of families are essentially the same as existing Pfam families. About the same proportion of families can find counterpart families in Pfam, but the boundary is fundamentally different. It could represent domain boundary extension, domain split or domain merge, reflecting how much ECOD domains help to improve domain boundary definition.

The most interesting part is the category of new families, which are as many as 3,266 when compared with Pfam (Fig. 4.3(d)) and 1,962 when compared with CDD. I have shown that most of them are small due to PDB bias, but is it because structural studies somehow use a lot of proteins from highly limited phylogenetic branches for whatever reasons, such as experimental consideration? I then checked the phylogenetic distribution of sequences in the alignment and assigned if the family is mainly from one kingdom or superkingdom. It turned out that the taxonomy distributions between all ECOD families and new families do not show much difference.

I further decided to explore the top ECOD homology groups where new families come from. Compared with the most populated H-groups overall (Cheng et al., 2014), the rank overlaps greatly with Helix-turn-helix domain and Immunoglobulin-like domain at top. However, domain of short lengths, such as “omega toxin-related” and “beta-beta-alpha zinc fingers” appear to be overrepresented. This could indicate specialized functions, but more likely implies that such domains demand a special method to compare their similarity better.

Additionally, most of them here are also very small in size, which suggests the alignment and scoring are highly biased by cysteines, because there is barely any conserved pattern in thin profiles or even singletons (Fig. 4.3(e)). The unrooted tree of families in omega toxin-related topology group is shown in Fig. 4.3(f) with “identical families” of Pfam families colored in blue. It seems that new ECOD families are often grouped with some known Pfam family, and the distance between families in the same clade are similar for both Pfam and ECOD families. It may imply that new families in ECOD are close relatives of known Pfam families or they could be actually from the same family given the limitation of current methodology. But at least they are consistent with Pfam family definition in this homologous group.

Lastly, I went over the list of large new domains against CDD which have more than 100 sequences to check and annotate them. Interestingly, some domains were described long time ago, but somehow were still not included in domain family databases, such as several glycoside hydrolases 109, 120 in CAZy (Lombard, Golaconda Ramulu, Drula, Coutinho, & Henrissat, 2014). There are also real novel families, for example, a glucuronoyl esterase (PDB: 4G4J) was a recently discovered enzyme with 411 members in its ECOD family (Charavgi, Dimarogona, Topakas, Christakopoulos, & Chrysina, 2013). For many others, the common situation is that homology can be clearly defined, but it does not have confident scores to be assigned to existing families.

Discussion

In this chapter, I described the procedure I used to build multiple sequence alignments and profiles based on ECOD domains. I also demonstrated that the quality of alignments is comparable to Pfam alignments. Profile-to-profile comparison results suggest that the domain boundaries of a large proportion of existing families were adjusted, presumably improved with ECOD domain definition. The comparison also discovered an unexpected number of new families. Investigation of these new families revealed that most of them have few sequences, likely to be a bias from PDB database. However, it is worth noting that there is another factor confounding the analysis, and sequence search in general, simply the shorter the domain length, the harder to get statistically significant scores. As in many cases of the domain boundary conflicts between ECOD and Pfam, ECOD usually further splits domains into individual evolutionary units (Cheng et al., 2015). Especially when a short domain is split out of a much longer domain, the score of the model for the long region, sometimes a whole protein, tends to be lower. For disulfide bond-rich domains and zinc fingers, it is intrinsically difficult since the sequence similarity signal is usually dominated by the cysteines and their spacing.

The aggregate HMM library and alignment file are available for download on the ECOD website together with distributable files of ECOD versions. A webpage is created for each family showing various alignments interactively using MSAViewer (Yachdav et al., 2016), taxonomy distribution of sequences, and relationship to other ECOD families, Pfam and CDD families. The family information page is also linked to the family level on the tree view page which displays the classification hierarchically.

Our pipeline can be used to continually create new families from unmapped domains in ECOD. Through periodical updates, it will not only help ECOD to group domains consistently and will also facilitate dedicated studies about specific families and protein annotations as a complementary resource to existing domain databases. I also expect it to serve as the basis to classify all protein sequences in ECOD.

Methods

For Pfam mapping analysis, HMMER 3.1b2 (Eddy, 2011) was used to assign Pfam families (version 28) to the ECOD nonredundant set with a E-value cutoff of 1e-3. Pfam assignments were made sequentially based on E-value, alignment overlap of 20 residues or less was allowed between subsequent assignments. The number of Pfam assignment on each ECOD representative domain was counted and coverage was calculated per residue. Pfam families that were assigned to ECOD domains from different H-groups in the set were also analyzed.

Domain sequences and structures were taken from ECOD version 178. All sequence clustering to reduce redundancy was done by CD-hit (Fu et al., 2012) at different identity level without length consideration. Pairwise structure alignments with Dali (Holm & Park, 2000), TM-align (Zhang & Skolnick, 2005) and FAST (Zhu & Weng, 2005) were run for representatives in the same family. The structure alignments generated from coordinates were also adjusted to match the domain sequences by adding gaps, as some residues may be incomplete and ignored by the programs. I used these alignments as custom constraints for PROMALS3D (Pei & Grishin, 2014) rather than let it to search for structure template on its own, since all domains to align have structures.

For large groups with at least 5 domains, the core was trimmed from both ends of the alignment until the first column that has more than 70% aligned residues. Structurally conserved indexes were predicted for each reside with both structural and sequence information, including secondary structure, carbon beta contacts and PSSM, conservation and gap fraction derived from PSI-BLAST (Altschul et al., 1997) results as well as secondary structure predicted by PSIPRED (McGuffin, Bryson, & Jones, 2000). The core of the alignment was conservatively cut at the column where any domain shows an index larger than 0.71, a threshold used on the structurally conserved region prediction server (I. K. Huang et al., 2013).

Protein sequences from Uniprot reference proteomes were downloaded in May 2017 containing 9,123 proteomes. Profile built with alignment of structural domains was searched against 80% redundancy reference proteome dataset using HMMSEARCH with an inclusion threshold of 1e-10 to construct the seed alignment. Then family HMM profile was built from seed alignment using reference columns deducted from alignment of structural domains. A full alignment was created by searching the family profile against the whole reference proteomes database with an inclusion threshold of 1e-3.

Pfam dataset for LGA computation was constructed by taking Pfam alignments with annotated PDB information. The sequence extracted from the mapped PDB and range was checked with sequence in the alignment and inconsistent sequence was disregarded. In some cases, different isoforms are recorded by PDB authors than the default Uniprot isoform Pfam uses. Then for each family, all pair of aligned sequences in the multiple sequence alignment were extracted, and aligned positions were converted to PDB index ranges for LGA to score (Zemla, 2003). Scores for all pairs were averaged, and average scores for each family were collected for comparison.

HHsearch (Soding, 2005) profiles of Pfam version 31 and CDD domains were built with default parameters from downloaded alignments. E-value of 1e-5 was used as threshold for hit acceptance. But hits up to probability 90% were considered for database crosslinking and used for automatic naming. Non-overlapping hits were accepted if the overlap is less than 10 residues or less than 10% length of accepted hits. To be labeled as identical family, the length of profile-to-profile alignment needs to be more than 80% or within 10 residues of either the length of query or hit.

Figures

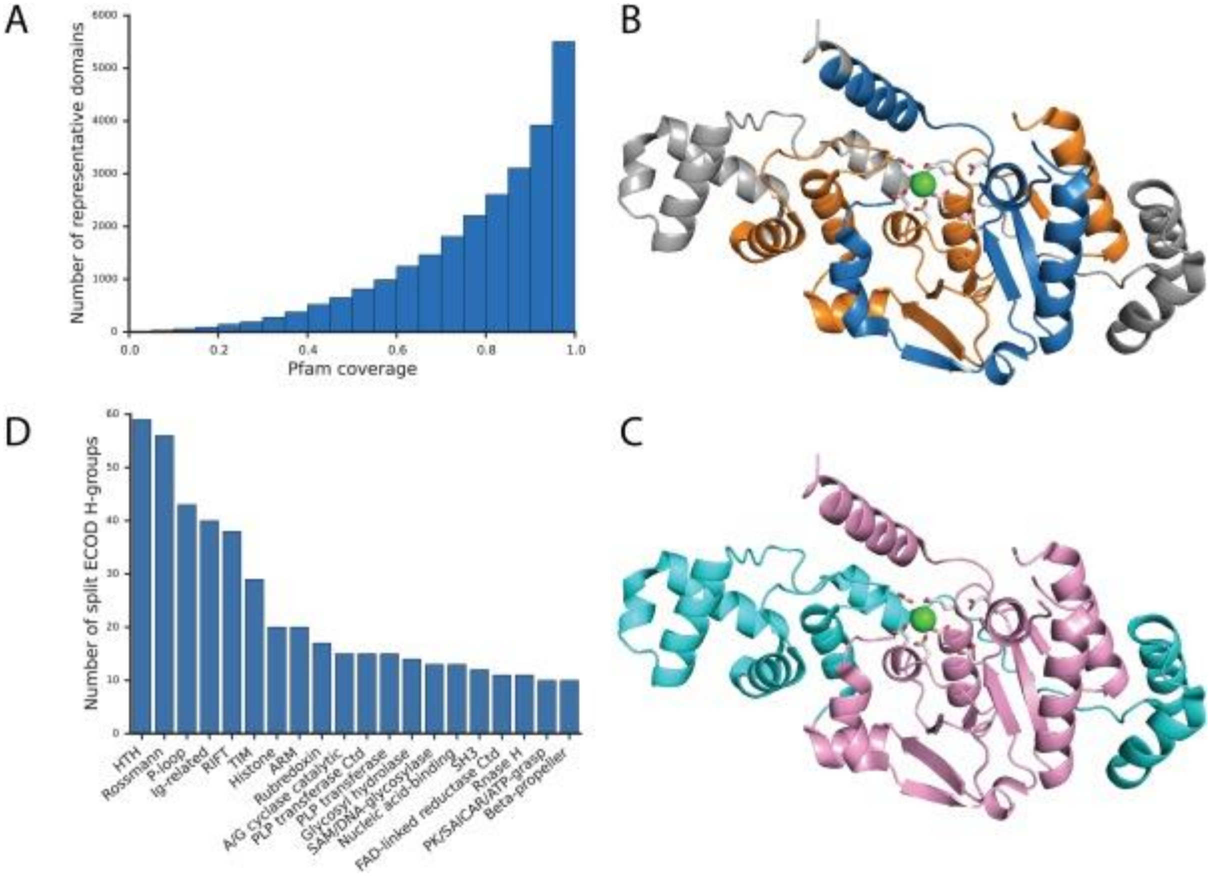


Fig. 4.1. Comparison of domain definitions between ECOD and Pfam.

(a) The distribution of Pfam family coverage on a nonredundant set of ECOD domains that have a one-to-one mapping to Pfam families. (b) Mapping of Pfam family XPG_N (PF00752, blue) and XPG_I (PF00867, orange) on RAD2 structure (PDB: 4q0w). (c) Mapping of HAD-related domain (e4q0wA2, pink) and SAM-like domain (e4q0wA1, cyan) from ECOD on the same structure. Side chains of catalytic residues are shown in stick, with the coordinating calcium ion in green sphere. (d) Top 20 H-groups where split Pfam families are assigned.

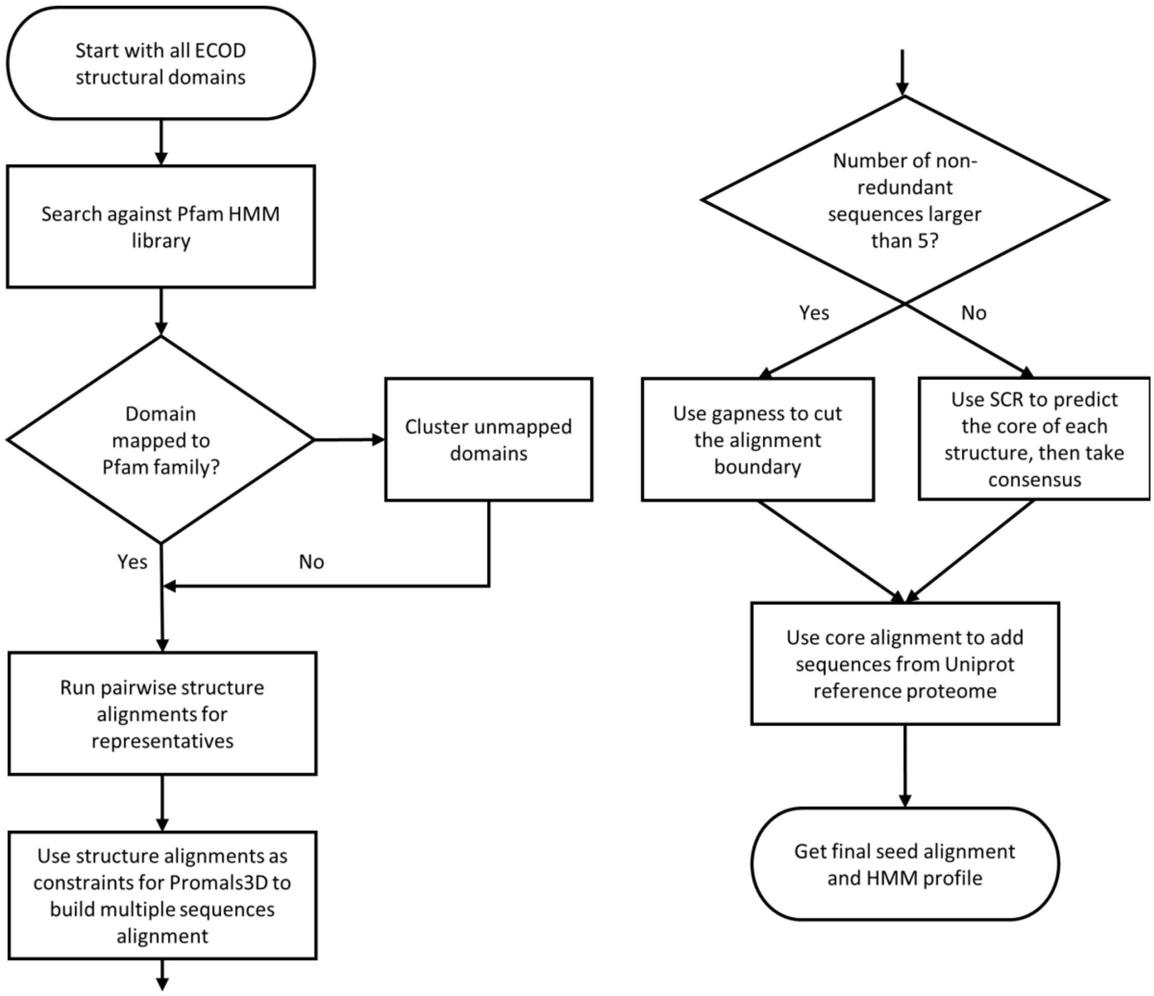
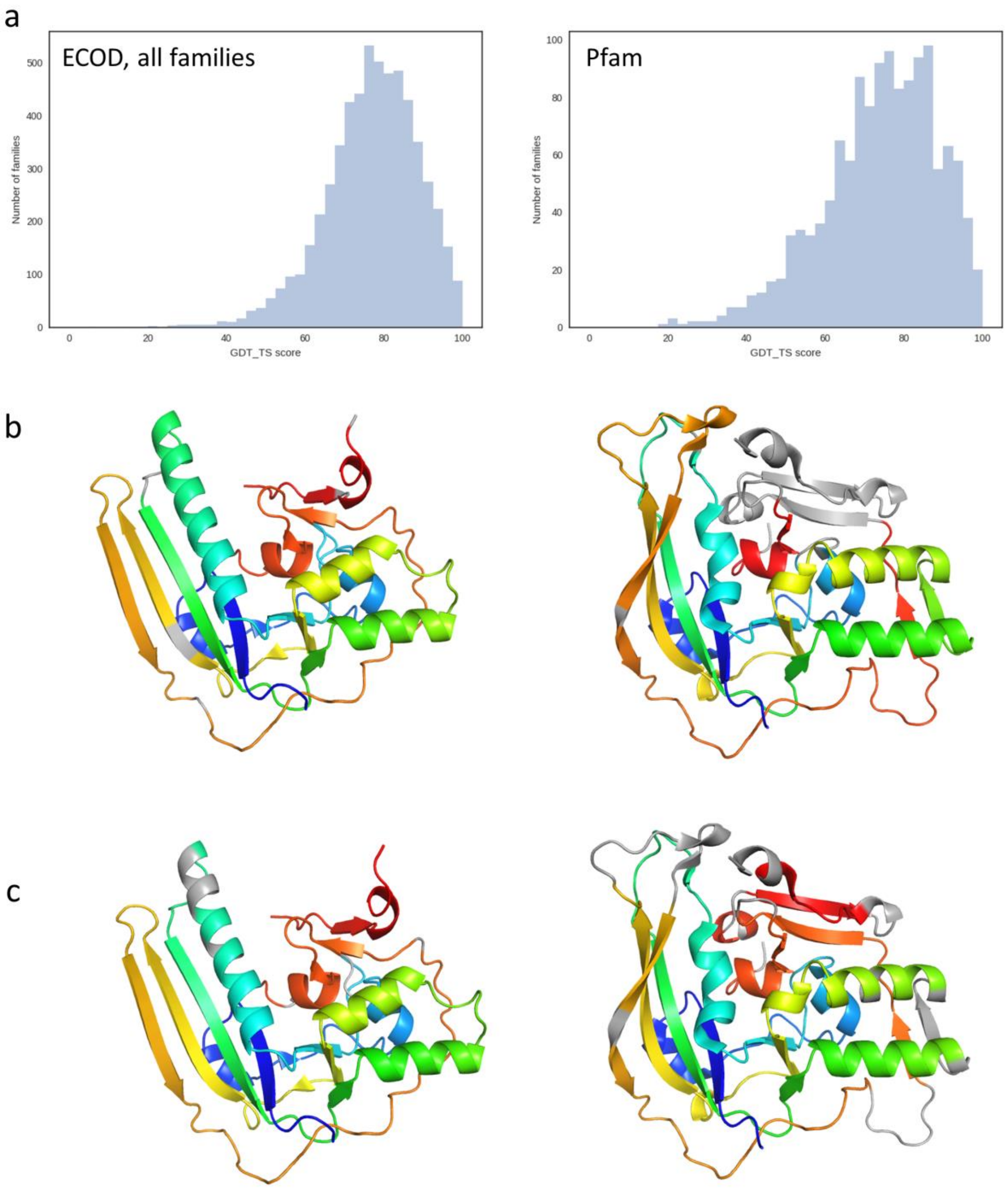

Start with all ECOD structural domains
Search against Pfam HMM library
Domain mapped to Pfam family?
Cluster unmapped domains
Yes
No
Run pairwise structure alignments for representatives
Use structure alignments as constraints for Promals3D to build multiple sequences alignment
Number of non-redundant sequences larger than 5?
Yes
No
Use gapness to cut the alignment boundary
Use SCR to predict the core of each structure, then take consensus
Use core alignment to add sequences from Uniprot reference proteome
Get final seed alignment and HMM profile

Fig. 4.2. Flowchart of the pipeline to build ECOD family alignment and profile.

Domains binned into the same Pfam or provisional families are collected and aligned by PROMALS3D with pairwise structure alignments as constraints and then possibly trimmed to the core region by consensus gapness or prediction. Then seed alignment and HMM profile is obtained by searching against Uniprot reference proteome with profile built from alignment of structural domains.

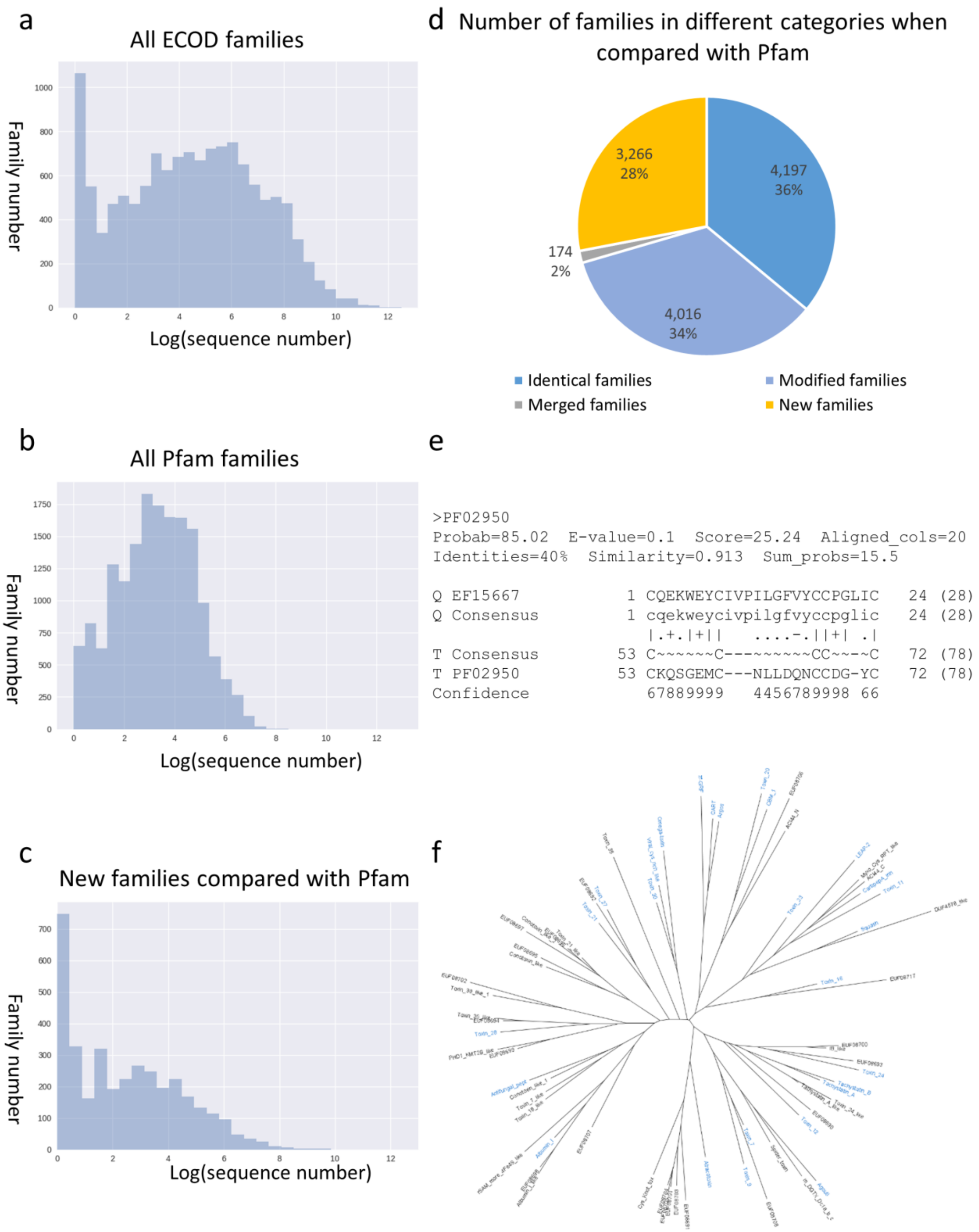


PDB: 1PTO
Pertussis toxin subunit 1

PDB: 4TLV
ADP-ribosylating toxin CARDS

Fig. 4.3. Validation of ECOD family alignment.

(a) The distribution of the average GDT_TS score per family of all ECOD families and Pfam families. (b) An example of Pfam alignment with registry shift mapped on the structures. Aligned residues are colored the same in rainbow from N-terminus. (c) ECOD family alignment of the same protein pair mapped on the structures. Proper gaps are made to handle loops and corresponding elements of different lengths.



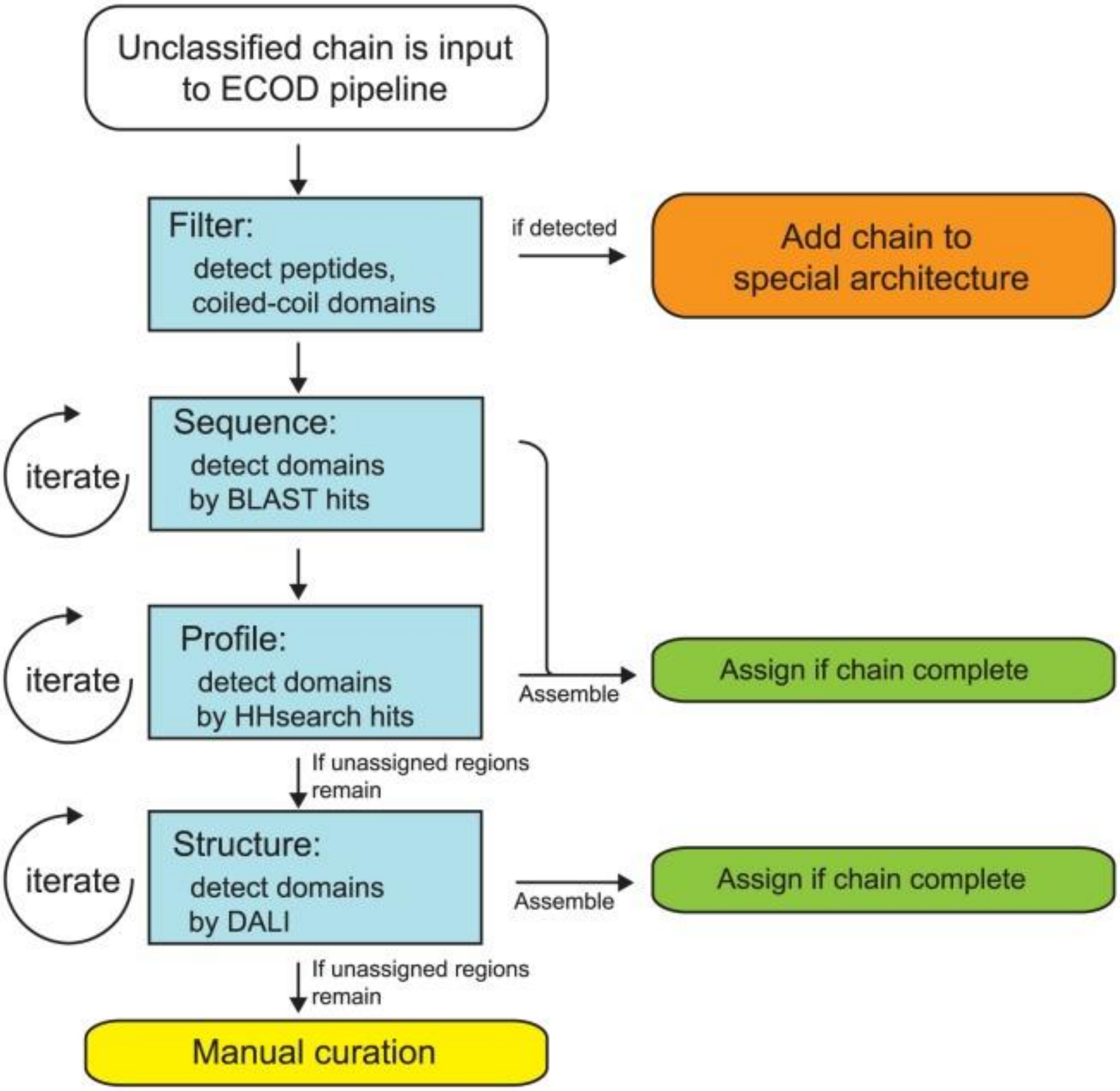
```
>PF02950
Probab=85.02  E-value=0.1  Score=25.24  Aligned_cols=20
Identities=40%  Similarity=0.913  Sum_probs=15.5

Q EF15667          1 CQEKWEYCIVPILGFVYCCPGLIC   24 (28)
Q Consensus        1 cqekweycivpilgfvyccpglic   24 (28)
                     |.+.|+||    ....-.||+|  .|
T Consensus       53 C~~~~~~C---~~~~~~~CC~~~C   72 (78)
T PF02950         53 CKQSGEMC---NLLDQNCCDG-YC   72 (78)
Confidence           67889999   4456789998 66
```

Fig. 4.4. Characterization of small families and new families in ECOD.

(a) The logarithm distribution of the number of sequences in ECOD families, showing a peak at very small size. (b) The logarithm distribution of the number of sequences in Pfam families for comparison. (c) The size distribution of only the families that cannot find a significant hit to Pfam by HHsearch. (d) The pie graph illustrates the proportion of four kinds of families when compared with Pfam. Identical family hits a Pfam family with comparable length. Modified family has a Pfam counterpart, but lengths differ substantially. Merged family has multiple non-overlapping Pfam hits. New family means no good Pfam hits. (e) An HHsearch alignment of an omega toxin family against Pfam as an example to show the difficulty to detect sequence similarity for small family, especially those domains with few secondary structure elements. Small family has a thin profile and does not exhibit too much conservation pattern. (f) An unrooted tree of all families in ECOD omega toxin-related topology group with identical families to Pfam colored in blue. New families are scattered and distributed with Pfam families, and the distances between families are comparable with distances between Pfam families.

APPENDIX A
Workflow of the ECOD automatic domain classification pipeline


Unclassified chain is input to ECOD pipeline
Filter: detect peptides, coiled-coil domains
if detected
Add chain to special architecture
iterate
Sequence: detect domains by BLAST hits
iterate
Profile: detect domains by HHsearch hits
Assemble
Assign if chain complete
If unassigned regions remain
iterate
Structure: detect domains by DALI
Assemble
Assign if chain complete
If unassigned regions remain
Manual curation

Unclassified structures enter from the top (white). Firstly, peptides, coiled-coils, and other unclassifiable regions are removed where possible and placed in their respective special architectures (orange). Secondly, unassigned regions of the input sequence are iteratively assigned by descending best hits from BLAST and HHsearch-based searches of ECOD databases. Assemblies of putative domains are optimized and assigned (green). If the chain is incomplete by sequence, a similar process occurs using DaliLite searches. If the query remains unclassified, it is manually curated (yellow).

BIBLIOGRAPHY

Abu-Safieh, L., Alrashed, M., Anazi, S., Alkuraya, H., Khan, A. O., Al-Owain, M., . . . Alkuraya, F. S. (2013). Autozygome-guided exome sequencing in retinal dystrophy patients reveals pathogenetic mutations and novel candidate disease genes. *Genome Res, 23*(2), 236-247. doi:10.1101/gr.144105.112

Alexandrov, N., & Shindyalov, I. (2003). PDP: protein domain parser. *Bioinformatics, 19*(3), 429-430.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res, 25*(17), 3389-3402.

Alva, V., Koretke, K. K., Coles, M., & Lupas, A. N. (2008). Cradle-loop barrels and the concept of metafolds in protein classification by natural descent. *Curr Opin Struct Biol, 18*(3), 358-365. doi:10.1016/j.sbi.2008.02.006

Andreeva, A., Howorth, D., Chothia, C., Kulesha, E., & Murzin, A. G. (2014). SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Res, 42*(Database issue), D310-314. doi:10.1093/nar/gkt1242

Aphasizhev, R., Aphasizheva, I., Nelson, R. E., & Simpson, L. (2003). A 100-kD complex of two RNA-binding proteins from mitochondria of Leishmania tarentolae catalyzes RNA annealing and interacts with several RNA editing components. *RNA, 9*(1), 62-76.

Arac, D., Boucard, A. A., Bolliger, M. F., Nguyen, J., Soltis, S. M., Sudhof, T. C., & Brunger, A. T. (2012). A novel evolutionarily conserved domain of cell-adhesion GPCRs mediates autoproteolysis. *EMBO J, 31*(6), 1364-1378. doi:10.1038/emboj.2012.26

Aravind, L., Anantharaman, V., Balaji, S., Babu, M. M., & Iyer, L. M. (2005). The many faces of the helix-turn-helix domain: transcription regulation and beyond. *FEMS Microbiol Rev, 29*(2), 231-262. doi:10.1016/j.femsre.2004.12.008

Bazan, J. F., & de Sauvage, F. J. (2009). Structural ties between cholesterol transport and morphogen signaling. *Cell, 138*(6), 1055-1056. doi:10.1016/j.cell.2009.09.006

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., . . . Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res, 28*(1), 235-242.

Bernardo, B. C., Belluoccio, D., Rowley, L., Little, C. B., Hansen, U., & Bateman, J. F. (2011). Cartilage intermediate layer protein 2 (CILP-2) is expressed in articular and meniscal cartilage and down-regulated in experimental osteoarthritis. *J Biol Chem, 286*(43), 37758-37767. doi:10.1074/jbc.M111.248039

Bjarnadottir, T. K., Fredriksson, R., Hoglund, P. J., Gloriam, D. E., Lagerstrom, M. C., & Schioth, H. B. (2004). The human and mouse repertoire of the adhesion family of G-protein-coupled receptors. *Genomics, 84*(1), 23-33. doi:10.1016/j.ygeno.2003.12.004

Buller, A. R., Freeman, M. F., Wright, N. T., Schildbach, J. F., & Townsend, C. A. (2012). Insights into cis-autoproteolysis reveal a reactive state formed through conformational rearrangement. *Proc Natl Acad Sci U S A, 109*(7), 2308-2313. doi:10.1073/pnas.1113633109

Burroughs, A. M., Allen, K. N., Dunaway-Mariano, D., & Aravind, L. (2006). Evolutionary genomics of the HAD superfamily: understanding the structural adaptations and catalytic diversity in a superfamily of phosphoesterases and allied enzymes. *J Mol Biol, 361*(5), 1003-1034. doi:10.1016/j.jmb.2006.06.049

Burroughs, A. M., Balaji, S., Iyer, L. M., & Aravind, L. (2007). Small but versatile: the extraordinary functional and structural diversity of the beta-grasp fold. *Biol Direct, 2*, 18. doi:10.1186/1745-6150-2-18

Busby, J. N., Panjikar, S., Landsberg, M. J., Hurst, M. R., & Lott, J. S. (2013). The BC component of ABC toxins is an RHS-repeat-containing protein encapsulation device. *Nature, 501*(7468), 547-550. doi:10.1038/nature12465

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics, 10*, 421. doi:10.1186/1471-2105-10-421

Capelson, M., Liang, Y., Schulte, R., Mair, W., Wagner, U., & Hetzer, M. W. (2010). Chromatin-bound nuclear pore components regulate gene expression in higher eukaryotes. *Cell, 140*(3), 372-383. doi:10.1016/j.cell.2009.12.054

Cappadocia, L., Marechal, A., Parent, J. S., Lepage, E., Sygusch, J., & Brisson, N. (2010). Crystal structures of DNA-Whirly complexes and their role in Arabidopsis organelle genome repair. *Plant Cell, 22*(6), 1849-1867. doi:10.1105/tpc.109.071399

Cappadocia, L., Parent, J. S., Zampini, E., Lepage, E., Sygusch, J., & Brisson, N. (2012). A conserved lysine residue of plant Whirly proteins is necessary for higher order protein assembly and protection against DNA damage. *Nucleic Acids Res, 40*(1), 258-269. doi:10.1093/nar/gkr740

Chai, J., Wu, J. W., Yan, N., Massague, J., Pavletich, N. P., & Shi, Y. (2003). Features of a Smad3 MH1-DNA complex. Roles of water and zinc in DNA binding. *J Biol Chem, 278*(22), 20327-20331. doi:10.1074/jbc.C300134200

Chandonia, J. M., Hon, G., Walker, N. S., Lo Conte, L., Koehl, P., Levitt, M., & Brenner, S. E. (2004). The ASTRAL Compendium in 2004. *Nucleic Acids Res, 32*(Database issue), D189-192. doi:10.1093/nar/gkh034

Charavgi, M. D., Dimarogona, M., Topakas, E., Christakopoulos, P., & Chrysina, E. D. (2013). The structure of a novel glucuronoyl esterase from Myceliophthora thermophila gives new insights into its role as a potential biocatalyst. *Acta Crystallogr D Biol Crystallogr, 69*(Pt 1), 63-73. doi:10.1107/S0907444912042400

Chaudhuri, I., Soding, J., & Lupas, A. N. (2008). Evolution of the beta-propeller fold. *Proteins, 71*(2), 795-803. doi:10.1002/prot.21764

Cheek, S., Qi, Y., Krishna, S. S., Kinch, L. N., & Grishin, N. V. (2004). 4SCOPmap: automated assignment of protein structures to evolutionary superfamilies. *BMC Bioinformatics, 5*, 197. doi:10.1186/1471-2105-5-197

Cheng, H., Kim, B. H., & Grishin, N. V. (2008). Discrimination between distant homologs and structural analogs: lessons from manually constructed, reliable data sets. *J Mol Biol, 377*(4), 1265-1278. doi:10.1016/j.jmb.2007.12.076

Cheng, H., Liao, Y., Schaeffer, R. D., & Grishin, N. V. (2015). Manual classification strategies in the ECOD database. *Proteins, 83*(7), 1238-1251. doi:10.1002/prot.24818

Cheng, H., Schaeffer, R. D., Liao, Y., Kinch, L. N., Pei, J., Shi, S., . . . Grishin, N. V. (2014). ECOD: an evolutionary classification of protein domains. *PLoS Comput Biol, 10*(12), e1003926. doi:10.1371/journal.pcbi.1003926

Coles, M., Djuranovic, S., Soding, J., Frickey, T., Koretke, K., Truffault, V., . . . Lupas, A. N. (2005). AbrB-like transcription factors assume a swapped hairpin fold that is evolutionarily related to double-psi beta barrels. *Structure, 13*(6), 919-928. doi:10.1016/j.str.2005.03.017

Copley, R. R., & Bork, P. (2000). Homology among (betaalpha)(8) barrels: implications for the evolution of metabolic pathways. *J Mol Biol, 303*(4), 627-641. doi:10.1006/jmbi.2000.4152

Crooks, G. E., Hon, G., Chandonia, J. M., & Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res, 14*(6), 1188-1190. doi:10.1101/gr.849004

D'Osualdo, A., Weichenberger, C. X., Wagner, R. N., Godzik, A., Wooley, J., & Reed, J. C. (2011). CARD8 and NLRP1 undergo autoproteolytic processing through a ZU5-like domain. *PLoS One, 6*(11), e27396. doi:10.1371/journal.pone.0027396

Desveaux, D., Allard, J., Brisson, N., & Sygusch, J. (2002). A new family of plant transcription factors displays a novel ssDNA-binding surface. *Nat Struct Biol, 9*(7), 512-517. doi:10.1038/nsb814

Desveaux, D., Subramaniam, R., Despres, C., Mess, J. N., Levesque, C., Fobert, P. R., . . . Brisson, N. (2004). A "Whirly" transcription factor is required for salicylic acid-dependent disease resistance in Arabidopsis. *Dev Cell, 6*(2), 229-240.

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics, 14*(9), 755-763.

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Comput Biol, 7*(10), e1002195. doi:10.1371/journal.pcbi.1002195

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res, 32*(5), 1792-1797. doi:10.1093/nar/gkh340

Finger, J. N., Lich, J. D., Dare, L. C., Cook, M. N., Brown, K. K., Duraiswami, C., . . . Gough, P. J. (2012). Autolytic proteolysis within the function to find domain (FIIND) is required for NLRP1 inflammasome activity. *J Biol Chem, 287*(30), 25030-25037. doi:10.1074/jbc.M112.378323

Finkelstein, A. V., & Ptitsyn, O. B. (1987). Why do globular proteins fit the limited set of folding patterns? *Prog Biophys Mol Biol, 50*(3), 171-190.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., . . . Punta, M. (2014). Pfam: the protein families database. *Nucleic Acids Res, 42*(Database issue), D222-230. doi:10.1093/nar/gkt1223

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res, 39*(Web Server issue), W29-37. doi:10.1093/nar/gkr367

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., . . . Bateman, A. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res, 44*(D1), D279-285. doi:10.1093/nar/gkv1344

Fontoura, B. M., Blobel, G., & Matunis, M. J. (1999). A conserved biogenesis pathway for nucleoporins: proteolytic processing of a 186-kilodalton precursor generates Nup98 and the novel nucleoporin, Nup96. *J Cell Biol, 144*(6), 1097-1112.

Fox, N. K., Brenner, S. E., & Chandonia, J. M. (2014). SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res, 42*(Database issue), D304-309. doi:10.1093/nar/gkt1240

Frickey, T., & Lupas, A. (2004). CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics, 20*(18), 3702-3704. doi:10.1093/bioinformatics/bth444

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics, 28*(23), 3150-3152. doi:10.1093/bioinformatics/bts565

Gough, S. M., Slape, C. I., & Aplan, P. D. (2011). NUP98 gene fusions and hematopoietic malignancies: common themes and new biologic insights. *Blood, 118*(24), 6247-6257. doi:10.1182/blood-2011-07-328880

Graebsch, A., Roche, S., Kostrewa, D., Soding, J., & Niessing, D. (2010). Of bits and bugs--on the use of bioinformatics and a bacterial crystal structure to solve a eukaryotic repeat-protein structure. *PLoS One, 5*(10), e13402. doi:10.1371/journal.pone.0013402

Graebsch, A., Roche, S., & Niessing, D. (2009). X-ray structure of Pur-alpha reveals a Whirly-like fold and an unusual nucleic-acid binding surface. *Proc Natl Acad Sci U S A, 106*(44), 18521-18526. doi:10.1073/pnas.0907990106

Griffis, E. R., Xu, S., & Powers, M. A. (2003). Nup98 localizes to both nuclear and cytoplasmic sides of the nuclear pore and binds to two distinct nucleoporin subcomplexes. *Mol Biol Cell, 14*(2), 600-610. doi:10.1091/mbc.E02-09-0582

Grishin, N. V. (2001a). Fold change in evolution of protein structures. *J Struct Biol, 134*(2-3), 167-185. doi:10.1006/jsbi.2001.4335

Grishin, N. V. (2001b). KH domain: one motif, two folds. *Nucleic Acids Res, 29*(3), 638-643.

Grishin, N. V. (2001c). Mh1 domain of Smad is a degraded homing endonuclease. *J Mol Biol, 307*(1), 31-37. doi:10.1006/jmbi.2000.4486

Hadley, C., & Jones, D. T. (1999). A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. *Structure, 7*(9), 1099-1112.

Haft, D. H., Selengut, J. D., Richter, R. A., Harkins, D., Basu, M. K., & Beck, E. (2013). TIGRFAMs and Genome Properties in 2013. *Nucleic Acids Res, 41*(Database issue), D387-395. doi:10.1093/nar/gks1234

Hanson, R. M., Prilusky, J., Renjian, Z., Nakane, T., & Sussman, J. L. (2013). JSmol and the Next-Generation Web-Based Representation of 3D Molecular Structure as Applied to Proteopedia. *Israel Journal of Chemistry, 53*(3-4), 207-216. doi:10.1002/ijch.201300024

Harrington, J. J., & Lieber, M. R. (1994). Functional domains within FEN-1 and RAD2 define a family of structure-specific endonucleases: implications for nucleotide excision repair. *Genes Dev, 8*(11), 1344-1355.

Heinz, L. X., Rebsamen, M., Rossi, D. C., Staehli, F., Schroder, K., Quadroni, M., . . . Tschopp, J. (2012). The death domain-containing protein Unc5CL is a novel MyD88-independent activator of the pro-inflammatory IRAK signaling cascade. *Cell Death Differ, 19*(4), 722-731. doi:10.1038/cdd.2011.147

Hodel, A. E., Hodel, M. R., Griffis, E. R., Hennig, K. A., Ratner, G. A., Xu, S., & Powers, M. A. (2002). The three-dimensional structure of the autoproteolytic, nuclear pore-targeting domain of the human nucleoporin Nup98. *Mol Cell, 10*(2), 347-358.

Holm, L., & Park, J. (2000). DaliLite workbench for protein structure comparison. *Bioinformatics, 16*(6), 566-567.

Holm, L., & Rosenstrom, P. (2010). Dali server: conservation mapping in 3D. *Nucleic Acids Res, 38*(Web Server issue), W545-549. doi:10.1093/nar/gkq366

Holm, L., & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J Mol Biol, 233*(1), 123-138. doi:10.1006/jmbi.1993.1489

Holm, L., & Sander, C. (1996). Mapping the protein universe. *Science, 273*(5275), 595-603.

Hu, Z., Yan, C., Liu, P., Huang, Z., Ma, R., Zhang, C., . . . Chai, J. (2013). Crystal structure of NLRC4 reveals its autoinhibition mechanism. *Science, 341*(6142), 172-175. doi:10.1126/science.1236381

Huang, I. K., Pei, J., & Grishin, N. V. (2013). Defining and predicting structurally conserved regions in protein superfamilies. *Bioinformatics, 29*(2), 175-181. doi:10.1093/bioinformatics/bts682

Huang, Y. S., Chiang, N. Y., Hu, C. H., Hsiao, C. C., Cheng, K. F., Tsai, W. P., . . . Lin, H. H. (2012). Activation of myeloid cell-specific adhesion class G protein-coupled receptor EMR2 via ligation-induced translocation and interaction of receptor subunits in lipid raft microdomains. *Mol Cell Biol, 32*(8), 1408-1420. doi:10.1128/MCB.06557-11

Huo, L., Wen, W., Wang, R., Kam, C., Xia, J., Feng, W., & Zhang, M. (2011). Cdc42-dependent formation of the ZO-1/MRCKbeta complex at the leading edge controls cell migration. *EMBO J, 30*(4), 665-678. doi:10.1038/emboj.2010.353

Huo, L., Zhang, H., Huo, X., Yang, Y., Li, X., & Yin, Y. (2017). pHMM-tree: phylogeny of profile hidden Markov models. *Bioinformatics, 33*(7), 1093-1095. doi:10.1093/bioinformatics/btw779

Hwang, K. Y., Baek, K., Kim, H. Y., & Cho, Y. (1998). The crystal structure of flap endonuclease-1 from Methanococcus jannaschii. *Nat Struct Biol, 5*(8), 707-713. doi:10.1038/1406

Ipsaro, J. J., & Mondragon, A. (2010). Structural basis for spectrin recognition by ankyrin. *Blood, 115*(20), 4093-4101. doi:10.1182/blood-2009-11-255604

Johnson, K., Farley, D., Hu, S. I., & Terkeltaub, R. (2003). One of two chondrocyte-expressed isoforms of cartilage intermediate-layer protein functions as an insulin-like growth factor 1 antagonist. *Arthritis Rheum, 48*(5), 1302-1314. doi:10.1002/art.10927

Kall, L., Krogh, A., & Sonnhammer, E. L. (2007). Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. *Nucleic Acids Res, 35*(Web Server issue), W429-432. doi:10.1093/nar/gkm256

Kalverda, B., Pickersgill, H., Shloma, V. V., & Fornerod, M. (2010). Nucleoporins directly stimulate expression of developmental and cell-cycle genes inside the nucleoplasm. *Cell, 140*(3), 360-371. doi:10.1016/j.cell.2010.01.011

Kim, B. H., Cheng, H., & Grishin, N. V. (2009). HorA web server to infer homology between proteins using sequence and structural similarity. *Nucleic Acids Res, 37*(Web Server issue), W532-538. doi:10.1093/nar/gkp328

Kim, Y. M., Stone, M., Hwang, T. H., Kim, Y. G., Dunlevy, J. R., Griffin, T. J., & Kim, D. H. (2012). SH3BP4 is a negative regulator of amino acid-Rag GTPase-mTORC1 signaling. *Mol Cell, 46*(6), 833-846. doi:10.1016/j.molcel.2012.04.007

Krasnoperov, V., Lu, Y., Buryanovsky, L., Neubert, T. A., Ichtchenko, K., & Petrenko, A. G. (2002). Post-translational proteolytic processing of the calcium-independent receptor of alpha-latrotoxin (CIRL), a natural chimera of the cell adhesion protein and the G protein-coupled receptor. Role of the G protein-coupled receptor proteolysis site (GPS) motif. *J Biol Chem, 277*(48), 46518-46526. doi:10.1074/jbc.M206415200

Krause, K., Kilbienski, I., Mulisch, M., Rodiger, A., Schafer, A., & Krupinska, K. (2005). DNA-binding proteins of the Whirly family in Arabidopsis thaliana are targeted to the organelles. *FEBS Lett, 579*(17), 3707-3712. doi:10.1016/j.febslet.2005.05.059

Krishna, S. S., & Grishin, N. V. (2004). Structurally analogous proteins do exist! *Structure, 12*(7), 1125-1127. doi:10.1016/j.str.2004.06.004

Krishnan, A., Almen, M. S., Fredriksson, R., & Schioth, H. B. (2012). The origin of GPCRs: identification of mammalian like Rhodopsin, Adhesion, Glutamate and Frizzled GPCRs in fungi. *PLoS One, 7*(1), e29817. doi:10.1371/journal.pone.0029817

Lagerstrom, M. C., & Schioth, H. B. (2008). Structural diversity of G protein-coupled receptors and significance for drug discovery. *Nat Rev Drug Discov, 7*(4), 339-357. doi:10.1038/nrd2518

Langenhan, T., Aust, G., & Hamann, J. (2013). Sticky signaling--adhesion class G protein-coupled receptors take the stage. *Sci Signal, 6*(276), re3. doi:10.1126/scisignal.2003825

Leonardo, E. D., Hinck, L., Masu, M., Keino-Masu, K., Ackerman, S. L., & Tessier-Lavigne, M. (1997). Vertebrate homologues of C. elegans UNC-5 are candidate netrin receptors. *Nature, 386*(6627), 833-838. doi:10.1038/386833a0

Liang, Y., Franks, T. M., Marchetto, M. C., Gage, F. H., & Hetzer, M. W. (2013). Dynamic association of NUP98 with the human genome. *PLoS Genet, 9*(2), e1003308. doi:10.1371/journal.pgen.1003308

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M., & Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res, 42*(Database issue), D490-495. doi:10.1093/nar/gkt1178

Lorenzo, P., Neame, P., Sommarin, Y., & Heinegard, D. (1998). Cloning and deduced amino acid sequence of a novel cartilage protein (CILP) identifies a proform including a nucleotide pyrophosphohydrolase. *J Biol Chem, 273*(36), 23469-23475.

Lupas, A., Engelhardt, H., Peters, J., Santarius, U., Volker, S., & Baumeister, W. (1994). Domain structure of the Acetogenium kivui surface layer revealed by electron crystallography and sequence analysis. *J Bacteriol, 176*(5), 1224-1233.

Lupas, A. N., Ponting, C. P., & Russell, R. B. (2001). On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *J Struct Biol, 134*(2-3), 191-203. doi:10.1006/jsbi.2001.4393

Macao, B., Johansson, D. G., Hansson, G. C., & Hard, T. (2006). Autoproteolysis coupled to protein folding in the SEA domain of the membrane-bound MUC1 mucin. *Nat Struct Mol Biol, 13*(1), 71-76. doi:10.1038/nsmb1035

Marchler-Bauer, A., & Bryant, S. H. (2004). CD-Search: protein domain annotations on the fly. *Nucleic Acids Res, 32*(Web Server issue), W327-331. doi:10.1093/nar/gkh454

Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R., Lu, S., Chitsaz, F., Geer, L. Y., . . . Bryant, S. H. (2015). CDD: NCBI's conserved domain database. *Nucleic Acids Res, 43*(Database issue), D222-226. doi:10.1093/nar/gku1221

Marchler-Bauer, A., Lu, S., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., DeWeese-Scott, C., . . . Bryant, S. H. (2011). CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res, 39*(Database issue), D225-229. doi:10.1093/nar/gkq1189

McGuffin, L. J., Bryson, K., & Jones, D. T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics, 16*(4), 404-405.

Mietus, M., Nowak, E., Jaciuk, M., Kustosz, P., Studnicka, J., & Nowotny, M. (2014). Crystal structure of the catalytic core of Rad2: insights into the mechanism of substrate binding. *Nucleic Acids Res, 42*(16), 10762-10775. doi:10.1093/nar/gku729

Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A., & Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res, 41*(12), e121. doi:10.1093/nar/gkt263

Murzin, A. G. (1998). How far divergent evolution goes in proteins. *Curr Opin Struct Biol, 8*(3), 380-387.

Murzin, A. G., Brenner, S. E., Hubbard, T., & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol, 247*(4), 536-540. doi:10.1006/jmbi.1995.0159

Nagano, N., Orengo, C. A., & Thornton, J. M. (2002). One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J Mol Biol, 321*(5), 741-765.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., & Thornton, J. M. (1997). CATH--a hierarchic classification of protein domain structures. *Structure, 5*(8), 1093-1108.

Orengo, C. A., Sillitoe, I., Reeves, G., & Pearl, F. M. (2001). Review: what can structural classifications reveal about protein evolution? *J Struct Biol, 134*(2-3), 145-165. doi:10.1006/jsbi.2001.4398

Park, H. H., Lo, Y. C., Lin, S. C., Wang, L., Yang, J. K., & Wu, H. (2007). The death domain superfamily in intracellular signaling of apoptosis and inflammation. *Annu Rev Immunol, 25*, 561-586. doi:10.1146/annurev.immunol.25.022106.141656

Pei, J., & Grishin, N. V. (2014). PROMALS3D: multiple protein sequence alignment enhanced with evolutionary and three-dimensional structural information. *Methods Mol Biol, 1079*, 263-271. doi:10.1007/978-1-62703-646-7_17

Pei, J., Kim, B. H., & Grishin, N. V. (2008). PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res, 36*(7), 2295-2300. doi:10.1093/nar/gkn072

Perler, F. B., Xu, M. Q., & Paulus, H. (1997). Protein splicing and autoproteolysis mechanisms. *Curr Opin Chem Biol, 1*(3), 292-299.

Petersen, T. N., Brunak, S., von Heijne, G., & Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods, 8*(10), 785-786. doi:10.1038/nmeth.1701

Pierleoni, A., Martelli, P. L., & Casadio, R. (2008). PredGPI: a GPI-anchor predictor. *BMC Bioinformatics, 9*, 392. doi:10.1186/1471-2105-9-392

Ponting, C. P., & Russell, R. B. (2000). Identification of distant homologues of fibroblast growth factors suggests a common ancestor for all beta-trefoil proteins. *J Mol Biol, 302*(5), 1041-1047. doi:10.1006/jmbi.2000.4087

Prikryl, J., Watkins, K. P., Friso, G., van Wijk, K. J., & Barkan, A. (2008). A member of the Whirly family is a multifunctional RNA- and DNA-binding protein that is essential for chloroplast biogenesis. *Nucleic Acids Res, 36*(16), 5152-5165. doi:10.1093/nar/gkn492

Promel, S., Frickenhaus, M., Hughes, S., Mestek, L., Staunton, D., Woollard, A., . . . Langenhan, T. (2012). The GPS motif is a molecular switch for bimodal activities of adhesion class G protein-coupled receptors. *Cell Rep, 2*(2), 321-331. doi:10.1016/j.celrep.2012.06.015

Remmert, M., Biegert, A., Hauser, A., & Soding, J. (2011). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods, 9*(2), 173-175. doi:10.1038/nmeth.1818

Rose, P. W., Prlic, A., Bi, C., Bluhm, W. F., Christie, C. H., Dutta, S., . . . Burley, S. K. (2015). The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic Acids Res, 43*(Database issue), D345-356. doi:10.1093/nar/gku1214

Rosenblum, J. S., & Blobel, G. (1999). Autoproteolysis in nucleoporin biogenesis. *Proc Natl Acad Sci U S A, 96*(20), 11370-11375.

Roy, A., Kucukural, A., & Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc, 5*(4), 725-738. doi:10.1038/nprot.2010.5

Sampathkumar, P., Ozyurt, S. A., Do, J., Bain, K. T., Dickey, M., Rodgers, L. A., . . . Burley, S. K. (2010). Structures of the autoproteolytic domain from the Saccharomyces cerevisiae nuclear pore complex component, Nup145. *Proteins, 78*(8), 1992-1998. doi:10.1002/prot.22707

Schaeffer, R. D., Kinch, L. N., Liao, Y., & Grishin, N. V. (2016). Classification of proteins with shared motifs and internal repeats in the ECOD database. *Protein Sci, 25*(7), 1188-1203. doi:10.1002/pro.2893

Schaeffer, R. D., Liao, Y., Cheng, H., & Grishin, N. V. (2017). ECOD: new developments in the evolutionary classification of domains. *Nucleic Acids Res, 45*(D1), D296-D302. doi:10.1093/nar/gkw1137

Scherly, D., Nouspikel, T., Corlet, J., Ucla, C., Bairoch, A., & Clarkson, S. G. (1993). Complementation of the DNA repair defect in xeroderma pigmentosum group G cells by a human cDNA related to yeast RAD2. *Nature, 363*(6425), 182-185. doi:10.1038/363182a0

Schrodinger, LLC. (2015). *The PyMOL Molecular Graphics System, Version 1.8*.

Schumacher, M. A., Karamooz, E., Zikova, A., Trantirek, L., & Lukes, J. (2006). Crystal structures of T. brucei MRP1/MRP2 guide-RNA binding complex reveal RNA matchmaking mechanism. *Cell, 126*(4), 701-711. doi:10.1016/j.cell.2006.06.047

Seki, S., Kawaguchi, Y., Chiba, K., Mikami, Y., Kizawa, H., Oya, T., . . . Ikegawa, S. (2005). A functional SNP in CILP, encoding cartilage intermediate layer protein, is associated with susceptibility to lumbar disc disease. *Nat Genet, 37*(6), 607-612. doi:10.1038/ng1557

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., . . . Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res, 13*(11), 2498-2504. doi:10.1101/gr.1239303

Sillitoe, I., Lewis, T. E., Cuff, A., Das, S., Ashford, P., Dawson, N. L., . . . Orengo, C. A. (2015). CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Res, 43*(Database issue), D376-381. doi:10.1093/nar/gku947

Silva, J. P., Lelianova, V., Hopkins, C., Volynski, K. E., & Ushkaryov, Y. (2009). Functional cross-interaction of the fragments produced by the cleavage of distinct adhesion G-protein-coupled receptors. *J Biol Chem, 284*(10), 6495-6506. doi:10.1074/jbc.M806979200

Soding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics, 21*(7), 951-960. doi:10.1093/bioinformatics/bti125

Soding, J., Biegert, A., & Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res, 33*(Web Server issue), W244-248. doi:10.1093/nar/gki408

Stein, U., Walther, W., Arlt, F., Schwabe, H., Smith, J., Fichtner, I., . . . Schlag, P. M. (2009). MACC1, a newly identified key regulator of HGF-MET signaling, predicts colon cancer metastasis. *Nat Med, 15*(1), 59-67. doi:10.1038/nm.1889

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., . . . Robinson, G. E. (2015). Big Data: Astronomical or Genomical? *PLoS Biol, 13*(7), e1002195. doi:10.1371/journal.pbio.1002195

Stuwe, T., von Borzyskowski, L. S., Davenport, A. M., & Hoelz, A. (2012). Molecular basis for the anchoring of proto-oncoprotein Nup98 to the cytoplasmic face of the nuclear pore complex. *J Mol Biol, 419*(5), 330-346. doi:10.1016/j.jmb.2012.03.024

Sun, Y., & Guo, H. C. (2008). Structural constraints on autoprocessing of the human nucleoporin Nup98. *Protein Sci, 17*(3), 494-505. doi:10.1110/ps.073311808

Tatusov, R. L., Koonin, E. V., & Lipman, D. J. (1997). A genomic perspective on protein families. *Science, 278*(5338), 631-637.

Teixeira, M. T., Siniossoglou, S., Podtelejnikov, S., Benichou, J. C., Mann, M., Dujon, B., . . . Fabre, E. (1997). Two functionally distinct domains generated by in vivo cleavage of Nup145p: a novel biogenesis pathway for nucleoporins. *EMBO J, 16*(16), 5086-5097. doi:10.1093/emboj/16.16.5086

The UniProt, C. (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Res, 45*(D1), D158-D169. doi:10.1093/nar/gkw1099

Thompson, J. D., Koehl, P., Ripp, R., & Poch, O. (2005). BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins, 61*(1), 127-136. doi:10.1002/prot.20527

Tinel, A., Janssens, S., Lippens, S., Cuenin, S., Logette, E., Jaccard, B., . . . Tschopp, J. (2007). Autoproteolysis of PIDD marks the bifurcation between pro-death caspase-2 and pro-survival NF-kappaB pathway. *EMBO J, 26*(1), 197-208. doi:10.1038/sj.emboj.7601473

Tomlinson, C. G., Atack, J. M., Chapados, B., Tainer, J. A., & Grasby, J. A. (2010). Substrate recognition and catalysis by flap endonucleases and related enzymes. *Biochem Soc Trans, 38*(2), 433-437. doi:10.1042/BST0380433

Tsutakawa, S. E., Classen, S., Chapados, B. R., Arvai, A. S., Finger, L. D., Guenther, G., . . . Tainer, J. A. (2011). Human flap endonuclease structures, DNA double-base flipping, and a unified understanding of the FEN1 superfamily. *Cell, 145*(2), 198-211. doi:10.1016/j.cell.2011.03.004

Van Walle, I., Lasters, I., & Wyns, L. (2005). SABmark--a benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics, 21*(7), 1267-1268. doi:10.1093/bioinformatics/bth493

Volynski, K. E., Silva, J. P., Lelianova, V. G., Atiqur Rahman, M., Hopkins, C., & Ushkaryov, Y. A. (2004). Latrophilin fragments behave as independent proteins that associate and signal on binding of LTX(N4C). *EMBO J, 23*(22), 4423-4433. doi:10.1038/sj.emboj.7600443

Vondruskova, E., van den Burg, J., Zikova, A., Ernst, N. L., Stuart, K., Benne, R., & Lukes, J. (2005). RNA interference analyses suggest a transcript-specific regulatory role for mitochondrial RNA-binding proteins MRP1 and MRP2 in RNA editing and other RNA processing in Trypanosoma brucei. *J Biol Chem, 280*(4), 2429-2438. doi:10.1074/jbc.M405933200

Wang, C., Yu, C., Ye, F., Wei, Z., & Zhang, M. (2012). Structure of the ZU5-ZU5-UPA-DD tandem of ankyrin-B reveals interaction surfaces necessary for ankyrin function. *Proc Natl Acad Sci U S A, 109*(13), 4822-4827. doi:10.1073/pnas.1200613109

Wang, R., Wei, Z., Jin, H., Wu, H., Yu, C., Wen, W., . . . Zhang, M. (2009). Autoinhibition of UNC5b revealed by the cytoplasmic domain structure of the receptor. *Mol Cell, 33*(6), 692-703. doi:10.1016/j.molcel.2009.02.016

Webb, B., & Sali, A. (2014). Protein structure modeling with MODELLER. *Methods Mol Biol, 1137*, 1-15. doi:10.1007/978-1-4939-0366-5_1

Wente, S. R., Rout, M. P., & Blobel, G. (1992). A new family of yeast nuclear pore complex proteins. *J Cell Biol, 119*(4), 705-723.

Yachdav, G., Wilzbach, S., Rauscher, B., Sheridan, R., Sillitoe, I., Procter, J., . . . Goldberg, T. (2016). MSAViewer: interactive JavaScript visualization of multiple sequence alignments. *Bioinformatics, 32*(22), 3501-3503. doi:10.1093/bioinformatics/btw474

Yoshida, K., Seo, H. S., Debler, E. W., Blobel, G., & Hoelz, A. (2011). Structural and functional analysis of an essential nucleoporin heterotrimer on the cytoplasmic face of the nuclear pore complex. *Proc Natl Acad Sci U S A, 108*(40), 16571-16576. doi:10.1073/pnas.1112846108

Zemla, A. (2003). LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res, 31*(13), 3370-3374.

Zhang, Y., & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res, 33*(7), 2302-2309. doi:10.1093/nar/gki524

Zhu, J., & Weng, Z. (2005). FAST: a novel protein structure alignment algorithm. *Proteins, 58*(3), 618-627. doi:10.1002/prot.20331