A PROPOSAL FOR THE PIRSF (PIR SUPERFAMILY) CLASSIFICATION SYSTEM

May 30, 2003
Protein Information Resource

Table of Content

A. Introduction – A Brief History

The concept of protein superfamilies was articulated by Margaret Dayhoff and the first comprehensive list that categorized all known complete or nearly complete protein sequences into superfamilies was published in 1976. At that time, 493 sequences were classified into 116 superfamilies. Sequences were further subdivided into families (closely related, >50% sequence identity), subfamilies (very closely related, >80% identity), and entries (nearly identical, >95% sequence identity).

The superfamily/family/subfamily organization was hierarchical and each protein or family belonged to one and only one superfamily. This fundamental principle was challenged by the discovery of multi-domain proteins. When genes for two enzymes fused, the resulting bifunctional protein might logically be placed into two superfamilies. Most scientists began to use labels such as "immunoglobulin superfamily" to refer to all proteins containing an immunoglobulin-related domain regardless of what other domains might be present. PIR responded to this dilemma in 1993 by distinguishing two types of superfamilies: domain superfamilies and homeomorphic superfamilies. Homeomorphic superfamilies contained proteins that are homologous from end to end and have the same domain architecture. By using the concept of homeomorphic protein superfamilies to organize all protein sequences, PIR could continue to place each protein sequence into one and only one protein superfamily and maintain the hierarchical organization. The homeomorphic superfamilies often consisted of one or a limited number of kinds of protein, e.g., plasminogen, and were convenient for propagation of annotation. For larger homeomorphic superfamilies, annotation could be propagated at the family level. However, the 50% sequence identity criterion was arbitrary and often did not partition a superfamily into natural groups for annotation.

A more serious problem has been the gap between the general usage of the terms "superfamily" and "family" in the scientific community and the specific usage of the term by PIR. In fact, PIR homeomorphic superfamilies correspond to what most scientists may call families. In relating PIR superfamilies to the other kinds of families catalogued in InterPro, our "superfamilies" are often children of groups that are called families. Now that PIR has joined SIB and EBI to produce a single protein sequence database (UniProt) and will continue to classify protein sequences into (super)families that are useful for annotation propagation, we have an urgent need to rethink our definitions of family and superfamily. At the same time, we will codify our practices (rules) for classification and spell out how we handle certain kinds of situations that arise from time to time.

B. Objectives and Definitions

In order to clearly define the PIR superfamily classification system, we should first determine the objective and scope of the classification system (PIRSF system hereafter). The PIRSF system is designed to facilitate the sensible propagation and standardization of protein annotation. Therefore:

1. To provide accurate and quality annotation, we will strive for excellence not completeness. Consequently, the PIRSF system will be manually curated – but with computer-assisted clustering.
2. To provide rich annotation, the PIRSF system will be integrated with other classification systems and specialized protein family databases for comprehensive information retrieval.
3. To annotate both biochemical and biological functions of proteins and to classify proteins without well-defined domains, the PIRSF will focus on classification of whole proteins, not individual domains.

Due to domain shuffling, a protein sequence may contain domains with different evolutionary histories. A hierarchical classification system cannot reflect this kind of relationship. Therefore, we define the PIRSF system as "**a network classification system based on evolutionary relationships of whole proteins.**" The primary nodes for curation are "**homeomorphic families,**" which consist of proteins that are both **homologous** (evolved from a common ancestor) and **homeomorphic** (sharing full-length sequence similarity and a common domain architecture). Above the "homeomorphic family" nodes in the network structure are parent "**superfamily**" nodes that connect a large number of distantly related families and orphan proteins based on common domains. They may be homeomorphic superfamilies, but are more likely to be "domain superfamilies" if the common domain regions do not extend to the entire full-length proteins. Below the "homeomorphic family" nodes are child "**subfamily**" nodes that are homologous and homoemorphic clusters representing functional specialization and/or domain architecture variation within a family.

Derived from the basic definition are the following three working definitions:

1. **Evolutionary relationship:** PIRSF members are **homologs** (may be orthologs or paralogs), as inferred by **detectable sequence similarity** [detailed in section E]. Members are assumed to have **common ancestry**. PIRSF families may range from those that are ancient and monophyletic (traceable to a Last Common Ancestor, or LCA) to those that are Lineage-Specific Expansions (LSEs).

2. **Homeomorphic:** Membership is based on the conceptual complete (full-length translation of the gene) sequence. Proteins are considered homeomorphic if they share **full-length sequence similarity and a common domain architecture**, as indicated by the same type, number, and order of defined domains for proteins within a defined length range [detailed in section E]. Length deviation may occur for alternative splice and alternate initiator variants, sequence fragments, and peptides derived from proteolytic processing. Variation of the domain architecture may exist for repeating domains and/or auxiliary domains. In contrast to evolutionarily conserved core domains present in all members, auxiliary domains are often mobile and may be easily lost, acquired, or functionally replaced during evolution.

3. **Network Structure:** PIRSF has a network structure with **parent-child relationships** to reflect the varying degrees of sequence conservation at different levels (superfamily, family, and subfamily). The threshold values may vary at each level depending on the

evolutionary rate in each group of proteins. Members of the child nodes are also members of the parent nodes. In the network, a homeomorphic family node may have multiple parent domain superfamily nodes, but a subfamily node has exactly one parent node.

C. Working Principles

Based on these definitions, we have derived the following working principles.

C.1. Curation Levels

1. **Multi-Level Curation:** There are three general curation levels, homeomorphic family and the levels above (superfamily) and below (subfamily). Multi-level classification is expected to improve protein annotation because multiple levels of sequence diversity allow for (i) more accurate extraction of conserved functional residues (for rule-based annotation), and (ii) classification of distantly related orphan proteins. The different curation levels are indicated by the parent-child relationship in a network structure depicted by a DAG (Direct Acyclic Graph) where each node represents a PIRSF superfamily, family, or subfamily. Each PIRSF curated node has a unique identifier (UID), which consists of the prefix "**PIRSF**" followed by six digits. Parent-child relationships are not evident from the UID.

2. **Homeomorphic Family Level:** Members in a homeomorphic family are homologous and homeomorphic ("**homeomorphic homologs**"). The homeomorphic family level is the primary PIRSF curation level – and most significant in terms of annotation and most invested with the biological meaning. A protein may be assigned to one and only one homeomorphic family, which may have zero or more parent nodes and zero or more child nodes. Curation at the homeomorphic family level includes: (i) mandatory text fields: family name, parent-child relationship, membership (member proteins), and signature domain architecture; and (ii) optional text fields: description, bibliography, and keyword/GO term. Each family also has a multiple sequence alignment, a phylogenetic tree, and full-length and domain HMMs, all of which are automatically generated from seed members. Families that are used to develop rules for propagating position-specific features (such as active/binding/catalytic sites) have manually curated multiple sequence alignments.

3. **Superfamily Level:** The superfamily level is used to bring together a number of distantly related families and orphan proteins that share one or more domains. Depending on the extent of domain coverage, a superfamily may be a "**homeomorphic superfamily**" (common domain architecture with full-length sequence coverage) or a "**domain superfamily**" (partial sequence coverage). Curation at the superfamily level includes: (i) mandatory text fields: superfamily name, parent-child relationship, membership (member families and orphan proteins), and common domain(s); (ii) optional text fields: description, bibliography, and keyword/GO term.

4. **Subfamily Level:** The subfamily level is used to delineate protein clusters within a homeomorphic family that have specialized functions and/or variable domain architectures. Like its parent, each subfamily is also homologous and homeomorphic. A protein may be assigned to zero or one subfamily, which will have exactly one parent node. Similar to curation at the homeomorphic family level, subfamily curation includes: (i) mandatory text fields: subfamily name, parent-child relationship, membership (member proteins), and signature domain architecture; and (ii) optional text fields: description, bibliography, and keyword/GO term. Each subfamily also has a multiple sequence alignment, a phylogenetic tree, and full-length and domain HMMs, as well as manually curated multiple sequence alignments as needed for position-specific rules.

C.2. Protein Annotation

The PIRSF system will be utilized to provide standardized and rich annotation for UniProt protein entries in three areas, namely PIRSF cross reference, position-specific feature annotation, and text annotation. Every curated PIRSF (at all three levels) will be incorporated into InterPro and cross-referenced in the UniProt "DR PIRSF" lines.

Position-specific features in UniProt "FT" (feature) lines will be annotated and propagated using a PIR rule-base system based on manually curated multiple sequence alignments and HMMs of homeomorphic families and subfamilies. Initial rules are being developed only for families/subfamilies that contain at least one member with known structure and experimentally determined site information. Evidence attribution includes a status tag ("experimental", "predicted", etc) and the rule itself (family information, multiple sequence alignments, HMMs, etc.).

Text annotation to be propagated (with evidence attribution) will include UniProt "DE" (description/protein name), "CC !-! Similarity" (family/domain name), "KW" (keyword), and "GO" (GO term) lines. Initial emphasis is on proteins in families/subfamilies that share common functions and contain sufficient numbers of experimentally verified members.

C.3. Integration with Other Classifications

PIRSF is cross-referenced to many other classification schemes, which are directly retrievable via iProClass. Domain-based retrieval can identify all PIRSFs sharing one or more Pfam domains. Likewise, CATH or SCOP-based retrieval can identify PIRSFs in common CATH homology levels or SCOP superfamily levels. In combination with the underlying taxonomic information, one can retrieve PIRSFs that occur only in given lineages or share common phyletic/phylogenetic profiles.

D. PIR Superfamily Redefined

D.1. What Are the Major Changes

1. The major change is that most of the groups that have been called homeomorphic superfamilies will now become homeomorphic families. Currently, there are 36,000 existing superfamilies, including 22,000 single-member superfamilies. Over 4000 superfamilies containing two or more members have been curated at the "first-tier" for their memberships and domain architectures. There are also 145,000 existing families that are automatically generated based on Fasta similarity search at a cut-off value of 50% sequence identity [Barker *et al*., 1996. Methods in Enzymology, 183, 59-71]. These existing superfamilies and families will now be referred to as superfamily and family clusters, and will provide a basis for defining new homeomorphic families [see section E.1].

2. We will use the term superfamily in much the way that most of the scientific community uses it to refer to all proteins related by some common structural feature (such as a particular kind of domain).

3. Proteins will no longer be constrained to belong to one and only one superfamily (but still in one and only one homeomorphic family).

4. It will be permissible to assign a protein to a superfamily without assigning it to a family (i.e. orphan proteins) because there is little return for curating single-member families.

D.2. What Will Stay the Same

1. PIRSF families (formerly homeomorphic superfamilies) will continue to be homeomorphic. The major PIR classification effort will continue to be at the level of homeomorphic families. A protein will belong to one and only one homeomorphic family.

2. A homeomorphic family will present evidence of common ancestry, traceable to a Last Common Ancestor (LCA) or to a Lineage-Specific Expansion (LSE).

D.3. What Will Be Modified or Enhanced

1. Partitioning a family into subfamilies will be optional and may be done to produce natural groupings for annotation.

2. There will not be an arbitrary level of similarity for defining either families or subfamilies. But there must be significant sequence similarity (sufficient to automatically generate an alignment of seed members) to group sequences into the same family.

3. In principle, parent-child relationships between families and subfamilies could be designated to any depth. In practice, family and subfamily should be more than enough for most situations. In principle, there could be additional parent-child relationships above the family level and below the domain superfamily level. In practice, we will seldom do that.

4. There will be additional caveats to the definition of homeomorphic. In selected families, auxiliary domains and/or repeats will be permitted to occur in varying numbers and sometimes to be absent, and some auxiliary domains may be permitted to vary in position (this will be rare).

E. Working Procedures/Implementations

E.1. Creating/Refining PIRSF Superfamily, Family and Subfamily

A systematic approach will be used to define related PIRSFs at all three levels in an iterative mode that couples manual curation with computer-assisted clustering and information retrieval. The steps described below will require further analysis and benchmarking to determine optimal parameters (marked as "TBD" for "to be determined") and to refine the process.

Step 1. Computer-Generated **Superfamily Clusters** Based on **Domains**: Retrieve all proteins sharing common domains and/or conserved regions in existing PIR superfamily (SF) and family (FAM) clusters [see section D.1.1] and their related orphan proteins (pre-computed and biweekly updated sequence neighbors at a threshold of e^{-5}).

Step 2. Computer-Generated **Homeomorphic Clusters** Based on **Full-Length Sequence Similarity**: (i) Filter proteins for sequence redundancy (stringent, 80-90% identity) [TBD] but preserve taxonomic diversity; (ii) Perform 'iterative Blastclust' to generate preliminary homeomorphic clusters. Blastclust is a single linkage clustering method with three parameters for score coverage, length coverage, and coverage on both neighbors. By fixing the length coverage (50-80%) [TBD] and neighbor coverage ("T" for "True"), iterative Blastclust can return preliminary homeomorphic clusters at different sequence similarity (score coverage) levels; (iii) Blast 'orphan proteins' (outside homeomorphic clusters) against the entire database to place additional members into the homeomorphic clusters based on top hits to existing clusters and protein length; (iv) Calculate the sequence distribution of each homeomorphic cluster, i.e., its homogeneity (distances amongst members) and uniqueness (distances to other clusters and orphan proteins) to measure the "goodness" of the cluster; (v) Retrieve member protein information, including protein name and pre-computed Pfam domain matches.

Step 3. Computer-Generated **Homeomorphic Subclusters** Based on **Sequence Similarity and Taxonomic Distribution**: (i) Filter proteins for sequence redundancy (less stringent, 50% identity, TBD) and for taxonomic distribution using a phyletic filter. The filter is used to derive a manageable set of representative sequences from large families and to assist in the analysis of phyletic patterns with common representatives; (ii) Perform iterative Blastclust and reciprocal Blast, calculate sequence distribution, and retrieve protein information as above.

Step 4. Curator-Defined **Homeomorphic Families** Based on **Sequence Similarity, Domain Architecture, and Taxonomic Distribution**: (i) Decide which Blastclust thresholds to

use for cluster definition. (ii) Further check the clusters by analyzing pre-computed reciprocal BLAST hit results. (iii) Perform multiple sequence alignments and neighbor-joining analysis on selected clusters/subclusters or sets of proteins and examine their domain architectures; (iv) Trace the homeomorphic cluster to a LCA (Last Common Ancestor) or a LSE (Lineage Specific Expansion); (v) Define the homeomorphic family and its parent-child relationships; (vi) Select seed members; (vii) Generate family-specific full-length and domain HMMs (automatically) based on seed members; (viii) Curate multiple sequence alignments if needed to propagate position-specific features.

Step 5. Curator-Defined **Homeomorphic Subfamilies** Based on **Domain and Functional Variation**: (i) Examine domain architecture variations among subclusters; (ii) Denote functional specializations among subclusters; (iii) Define homeomorphic subfamilies and their parent-child relationships; (iv) Select seed members, generate subfamily-specific HMMs, and curate sequence alignments if needed.

Step 6. Curator-Defined **Superfamilies** Based on Domains and Homeomorphic Families: (i) Retrieve all homeomorphic families containing the given domains and define their parent-child relationships with the superfamily; (ii) Identify distantly related orphan proteins (not classifiable into the homeomorphic families) using any one or all of the following: domain HMMs, PSI-Blast, or other program, based on homeomorphic family members.

E.2. Classifying New Members into Existing Families and Subfamilies

Procedures are also being developed and benchmarked for the recruitment of new members to existing PIRSF families and subfamilies. (Placing new orphan proteins into superfamilies is described above in step 6). New members not classified in the initial definition phase [section E.1] will fall into two major categories: new protein sequences entering the UniProt database, and related sequences not classified due to length deviation. To accommodate the length deviation, we need two types of membership: "**regular members**" for proteins sharing end-to-end sequence similarity and common domain architecture and "**associate members**." Members whose lengths are outside the family length range, including sequences fragments, alternate splice and alternate initiator variants, and peptides derived from proteolytic processing, are classified as associate members with the conceptual complete sequence from which they are derived. Associate members also include individual proteins with atypical domain architecture (thus, not yet forming a separate subfamily). Accordingly, we have separate procedures for placing new regular members and associate members.

F. Case Example – PIRSF001969

The following case example using insulin-like growth factor binding proteins (IGFBPs) illustrates when we can use PIRSF homeomorphic family and subfamily levels to improve classification and assist functional annotation.

F.1. Sequence Conservation and Evolutionary Relationship

We will derive a new **homeomorphic family** (PIRSF001969) based on the current superfamily (SF001969) for insulin-like growth factor binding proteins (IGFBPs). They are a group of secreted proteins, ranging from 227 to 328 amino acids long, which bind to IGF-I and IGF-II with high affinity and modulate the biological actions of IGFs. The family consists of 61 regular (non-fragment) members in PIR-NREF (containing non-redundant PIR, Swiss-Prot, TrEMBL, RefSeq, and GenPept sequences), representing a variety of vertebrate species ranging from mammals to fish. All family members share a common domain architecture (PF00086:PF00219), containing an N-terminal IGF binding protein domain (PF00086) and a C-terminal thyroglobulin type-1 repeat domain (PF00219).

While the N- and C-terminal domains in the IGFBP family are conserved, the mid-region (L region) is highly variable with protease cleavage sites and phosphorylation and glycosylation sites. According to the consensus nomenclature adopted by researchers in the field, the IGFBP family has six member types, IGFBP-1 through 6, based on their conserved intron/exon organization, sequence similarity, and high binding affinity to IGFs. These subgroups correspond directly to six automatically clustered PIR families within SF001969, namely FAM0001775, FAM0007601, FAM0001776, FAM0012205, FAM0012206, and FAM0033225. They also correspond exactly to clusters that are automatically generated by iterative Blastclust at different sequence similarity threshold values, and will become the basis for **six homeomorphic subfamilies**.

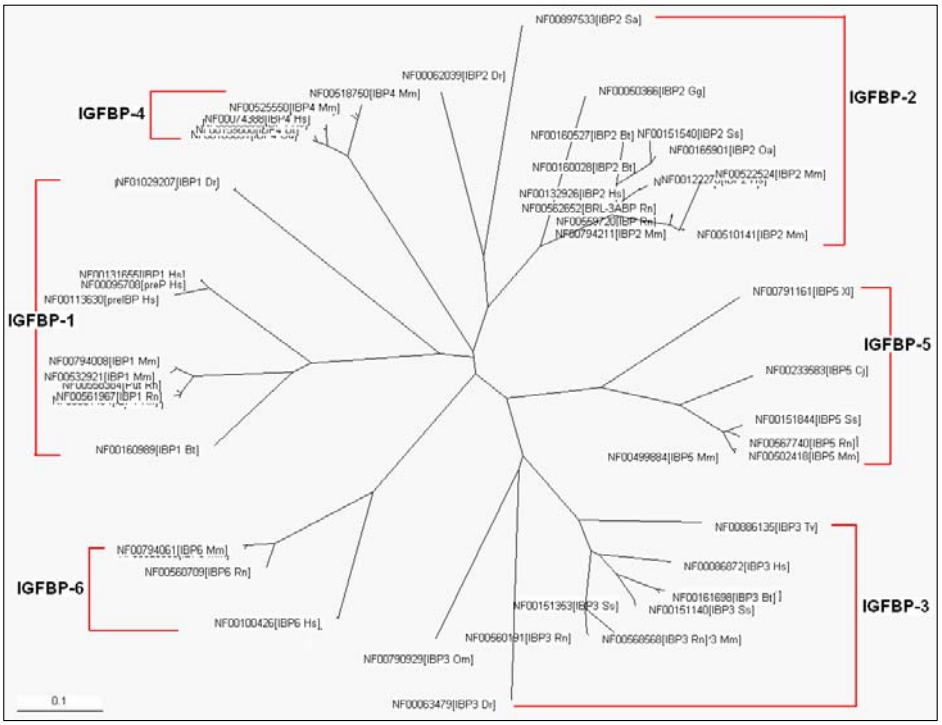

Figure 1.

The phylogenetic tree of this family (Figure 1) has six clearly distinguishable branches, corresponding to subfamilies for IGFBP-1 through -6. ClustalW was used to generate the multiple sequence alignment and the neighbor-joining tree (displayed in TreeView). Each of the

six nodes is supported by bootstrap values of 94% and above. Each subfamily contains sequences from various vertebrate species, indicating that they are orthologous sequences within the group.

IGFBP-1: Hs,Mm,Rn,Bt,Oa,Ss,Dr,
IGFBP-2: Hs,Mm,Rn,Bt,Oa,Ss,Sa,Dr,Gg,Cc
IGFBP-3: Hs,Mm,Rn,Bt,Oa,Ss,Om,Dr,Tv,Bb
IGFBP-4: Hs,Mm,Rn,Bt,Oa
IGFBP-5: Hs,Mm,Rn,Bt,Ss,Cj,Xl
IGFBP-6: Hs,Mm,Rn,Bt,Oa

Homo sapiens (Hs), Mus musculus (Mm), Rattus norvegicus (Rn), Bos taurus (Bt), Ovis aries (Oa), Sus scrofa (Ss), Bubalus bubalis (Bb), Trichosurus vulpecula (Tv), Gallus gallus (Gg), Coturnix japonica (Cj), Coturnix coturnix (Cc), Xenopus laevis (Xl), Danio rerio (Dr), Oreochromis mossambicus (Om), Sparus aurata (Sa)

F.2. Functional Diversity

Associated with the sequence variation among subfamilies is the functional diversity of the family. IGFBPs are unusually pleiotropic molecules with functions ranging from the traditional role of prolonging the half-life of the IGFs to functioning as growth factors independent of the IGFs. Examples of **IGF-independent functions** supported by experimental evidence include: IGFBP-5, which stimulates markers of bone formation in osteoblasts lacking functional IGFs, and IGFBP-1, which stimulates cell migration through integrin-mediated action. In addition, IGFBP-3 and -5 may also exert transcriptional activation of genes based on their nuclear localization.

A few conserved functional sites have been mapped to the IGFBP family and subfamilies.

1. IGF binding site, determined based on the 3D structure of an IGFBP-5 protein, is conserved among most family members.
2. Disulfide bonds are conserved in all IGFBP family members, except that IGFBP-6 has eight disulfide bonds, instead of nine as in IGFBP-1 through 5.
3. Integrin binding site is conserved in IGFBP-1 (with experimental evidence) and IGFBP-2.
4. A highly basic C-terminal region is conserved in IGFBP-3 and IGFBP-5, which contains an extra cellular matrix binding site and a putative nuclear localization signal.

This case serves as an example of how splitting homeomorphic families into subfamilies can assist in functional annotation. Conserved sites associated with functional information derived from experimentally validated members can be propagated to other family or subfamily members.