Proposal for the Definition of "Protein Superfamily"

SFDEF-0893 August 18, 1993 David G. George Protein Information Resource (PIR) National Biomedical Research Foundation (NBRF) Washington, DC

Margaret O. Dayhoff introduced the term **protein superfamily** in 1974 [1,2,3]. Since that time, the sequences in the PIR-International Protein Sequence Database have been classified into protein superfamilies. Prior to about 1990, the superfamily classification permitted a sequence to be assigned to a single superfamily only. The recognition of mosaic, multidomain proteins, whose component domains appear to have had separate evolutionary histories, has made this approach no longer effective. Moreover, the term **superfamily** has come into common usage and its meaning is no longer well defined. Although originally defined as a group of evolutionarily related proteins, it also has been used in the published literature to refer to a group of structurally or functionally related proteins not necessarily of common evolutionary origin.

The term domain has been employed in the Protein Sequence Database to mean a region of special biological interest within a single protein chain. This term also has been used with many different meanings; in particular, it has been used to characterize a region within the three-dimensional structure of a protein that may encompass regions of several distinct protein chains.

We have developed a conceptual model that allows all of these various meanings to be associated with the terms superfamily and domain. This model supports a protein superfamily classification scheme that partitions the Protein Sequence Database fully into superfamilies that are closed under transitivity, i.e., if A is in the same superfamily as B and B is in the same superfamily as C, then A and C are in the same superfamily for all A, B, and C in the superfamily.

Classification is the systematic arrangement into groups or categories according to established criteria [4]. The development of a classification scheme requires an abstraction on the data. "An abstraction is a mental process that we use when we select some characteristics and properties of a set of objects and exclude other characteristics that are not relevant. In other words, we apply an abstraction whenever we concentrate on properties of a set of objects that we regard as essential, and forget about their differences. ... The classification abstraction is used for defining one concept as a class of real-world objects characterized by common properties [5]." Because the design of a classification system requires abstraction, its utility cannot be assessed based on first principles. There is no one best classification system. The value of a classification scheme can be assessed only on pragmatic criteria, such as how well it characterizes the data, what insights can be gained by contrasting the properties of objects assigned to the same class, and the ease of clearly making class assignments.

The strategy employed here is to define a protein superfamily as an abstract class of protein elements. Superfamilies are composed of families that are also classes of protein elements; a superfamily is the union of the families from which it is composed. Superfamilies are characterized by the type of relationship that defines the class. We are concerned here specifically with evolutionary relationships among protein sequences. The problems associated with mosaic proteins are addressed by classifying protein sequence domains only. Sequence domains are defined as subsequences and are considered to be distinct when they correspond to different subsequences, even when the subsequences overlap or when one domain is contained within another. These overlap relationships cannot be ignored; however, they are orthogonal to the superfamily classification and are best treated as an external mappings among the domain classes, independent of the superfamily classification. These relationships will not be treated here.

The formulation presented here allows the Protein Sequence Database to be partitioned completely into homologous sequence domains. A special class of sequence domains whose members correspond to entire protein sequences (improper subsequences) is distinguished. We coin the term homeomorphic class to refer to such a class. We demonstrate that all sequences in the Protein Sequence Database can be assigned to homeomorphic superfamilies. The sequences in the database can be ordered by homeomorphic superfamily. This paper presents a set of self-consistent definitions that provide a formal conceptual architecture for the representation of family and superfamily classes within the Protein Sequence Database.

Preliminary Definitions and Characteristics of the Data

Definition: A macromolecular chain is a polymeric chain composed of an ordered set of molecular units linked by covalent bonds; these molecular units are called residues.

We use the term macromolecular chain exclusively to refer to unbranched macromolecular chains. Unless otherwise stated, this term is also used exclusively to refer to biological macromolecules.

Definition: A macromolecular sequence is a symbolic representation of the chemical structure of a macromolecular chain. It depicts the sequential ordering of residues in the chain.

Macromolecular sequences are represented in the Protein Sequence Database as sequences of characters. Such sequences are defined over a characteristic alphabet of valid residue symbols; the alphabet is specific to each type of macromolecule. Residues in a sequence are numbered starting at 1; these residue numbers are referred to as sequence positions. *Definition*: A protein is a macromolecule consisting of one or more polypeptide chains and possibly of other organic or inorganic conjugated chemical groups.

Definition: A polypeptide chain is a macromolecular chain consisting of amino acid residues linked from amino end to carboxyl end by peptide bonds.

Definition: A protein sequence is a symbolic representation of the chemical structure of a polypeptide chain.

We will use protein chain or simply chain (when the context is clear) as synonymous with polypeptide chain. Chains linked by bonds other than peptide bonds are considered to be separate chains even when the resulting molecule is unbranched. In the remainder of this paper, we will use the terms sequence and chain to refer exclusively to protein sequences and protein chains, although many of the following definitions are common to all macromolecular sequences. Moreover, the discussion will be restricted to protein sequences as represented in the PIR-International Protein Sequence Database. The Protein Sequence Database represents sequence data from naturally occurring, wild-type proteins only.

Definition: Sequence A is said to be a subsequence of Sequence B if all of the elements of A are contained in B in the same order. Subsequence A is said to be a contiguous subsequence of B if the elements of A are contiguous within B.

Definition: Sequence A is said to be equal to Sequence B if A is a subsequence of B and B is a subsequence of A.

Definition: Sequence A is said to be a proper subsequence of Sequence B if it is a subsequence of B, it is not the empty sequence, and it is not equal to B.

Although a macromolecular chain consists of a series of contiguous residues, a macromolecular sequence is permitted to be noncontiguous when it is incomplete. When this situation occurs it is understood that the chemical structure of the macromolecular chain has not been fully elucidated.

Definition: A complete sequence is a sequence that contains a residue symbol corresponding to each residue in the macromolecular chain that it symbolically represents.

Definition: A conceptual complete sequence is a conceptual image of a macromolecular chain. When the sequence of the macromolecular chain is not known completely, the corresponding conceptual complete sequence is not a complete sequence, i.e., the undetermined segments cannot be physically represented.

Definition: A sequence fragment is a contiguous, proper subsequence of a conceptual complete sequence.

All sequences in the Protein Sequence Database represent conceptual complete sequences. When the complete sequence is not known, the sequence is represented by an ordered set of fragments separated by break symbols; such sequences are referred to as incomplete sequences or fragments. The break symbol, the slash character, is a special residue symbol that indicates the absence of one or more residues at the indicated position. A single break symbol is used at the junction of two fragments and at the amino (left) and carboxyl (right) ends of the incomplete sequences when these terminal regions are absent. Note that the break symbol may also be used when a segment of a sequence has been deliberately omitted from the display even though the sequence of the omitted segment is known.

In addition to the break symbol, two other special residue symbols are included in the alphabet: (1) the termination symbol, an asterisk, which represents the terminal (right) end of the sequence (it is included at the end of all sequences); (2) the gap (or NUL) symbol, a dash character, which represents the absence of a residue at the indicated position in an alignment. The gap character is used in sequence alignments to indicate the absence of a residue at that site in the alignment. It is distinct from the break character in that it indicates that the sequence is complete at that position in the sequence but there is no equivalent residue at that position in the alignment.

Definition: A sequence alignment expresses a residue-by-residue relationship between macromolecular chains (or portions of the chains). This relationship is depicted by a matrix. Each row contains a sequence or subsequence. The columns contain the corresponding residues from each sequence. Gap symbols are inserted in the sequences or subsequences to maintain the position-wise relationship between corresponding residue symbols. By convention, incomplete sequences may be included on alignments. Missing segments, bordered by noncontiguities (break symbols), are indicated in the alignment by white space. No relationship is specified between the undetermined (missing) segments and the corresponding segments of other sequences present in the alignment; however, it is assumed that the missing segment would satisfy the relationship implied by the alignment and could be placed on the alignment if known. For example,

```
DEBOXA    3-Methyl-2-oxobutanoate dehydrogenase (lipoamide) (EC 1.2.4.4)
          alpha chain precursor - Bovine (fragments)
DEHUXA    3-Methyl-2-oxobutanoate dehydrogenase (lipoamide) (EC 1.2.4.4)
          alpha chain precursor - Human (fragments)
DERTXA    3-Methyl-2-oxobutanoate dehydrogenase (lipoamide) (EC 1.2.4.4)
          alpha chain    precursor - Rat (fragments)
DEBOXA  M Q G S A K M A M A V A V A V A R V W R P/H P H R W Q Q Q Q H F S S L/
DEHUXA                  / G G A I A A A R V W R L/H P P R - - Q Q Q Q F S S L/
DERTXA                        / S A A K I W R P/H P S R - Q Q Q Q Q F P S L/
```

depicts an alignment of several lipoamide dehydrogenase fragments. The `/' character delimits fragments. Dashes, `-', indicate gaps.

An alignment is a graphic representation of the hypothesis that the corresponding residues at each position along the alignment serve equivalent (as defined by the relationship) roles in each sequence or subsequence in the alignment. Alignments may be used to depict any relationship among the residues of protein chains provided that the relationship is a residue-by-residue mapping. For example, if the alignment

depicts a functional relationship between the chains, the alignment implies that each aligned residue serves an equivalent functional role in each of the represented chains.

Although alignments can be used to express compression-expansion events as well as deletion-insertion events [6,7],*alignments are interpreted here to depict residue-by-residue mappings only.

*Compression-expansion events would imply that the subalignment of WQQ with --Q and -QQ should be interpreted to mean that residues WQQ of sequence DEBOXA fulfill the same role as residue Q of sequence DEHUXA and residues QQ of sequence DERTXA. In contrast, deletion-insertion events would imply that W is unrelated to any residue in sequences DEHUXA and DERTXA; while the first Q from DEBOXA is related to the first Q from sequence DERTXA, sequence DEHUXA has no counterpart.

Sequence alignments depicting two sequences or subsequences are called pairwise sequence alignments; those depicting more than two sequences or subsequences are called multiple sequence alignments.

Definition: A global alignment is an alignment that represents complete macromolecular chains. Global alignments may contain incomplete sequences provided that they correspond to conceptual complete sequences.

Definition: A local alignment contains proper subsequences only and represents a relationship between subregions of macromolecular chains.

Families and Superfamilies

A class is a group, set, or kind sharing common attributes [4]. The term protein class will be used to refer to a group of protein sequences or subsequences sharing common attributes.

A set is said to be partitioned when it is separated into subsets such that every element is a member of some subset and no element is a member of more than one subset [4].

Definition: A protein family is a protein class that is distinguished as exhibiting a threshold level of some relationship. The relationship must allow the set of all protein sequences or subsequences to be partitioned and each family must be closed under transitivity. A family may contain a single member (under all relationships considered here a sequence is related to itself).

Definition: A protein superfamily is a protein class composed of one or more protein families; a superfamily is the union of its constituent families. The set of all superfamilies must constitute a partitioning of the set of all protein sequences or subsequences under the relationship that defines the protein families and each superfamily must be closed under transitivity. A superfamily may contain a single member.

Domains

Definition: A protein domain is a region within a protein that has been distinguished by a well-defined set of properties or characteristics.

Definition: A domain class is a class of domains that share a common set of well-defined properties or characteristics.

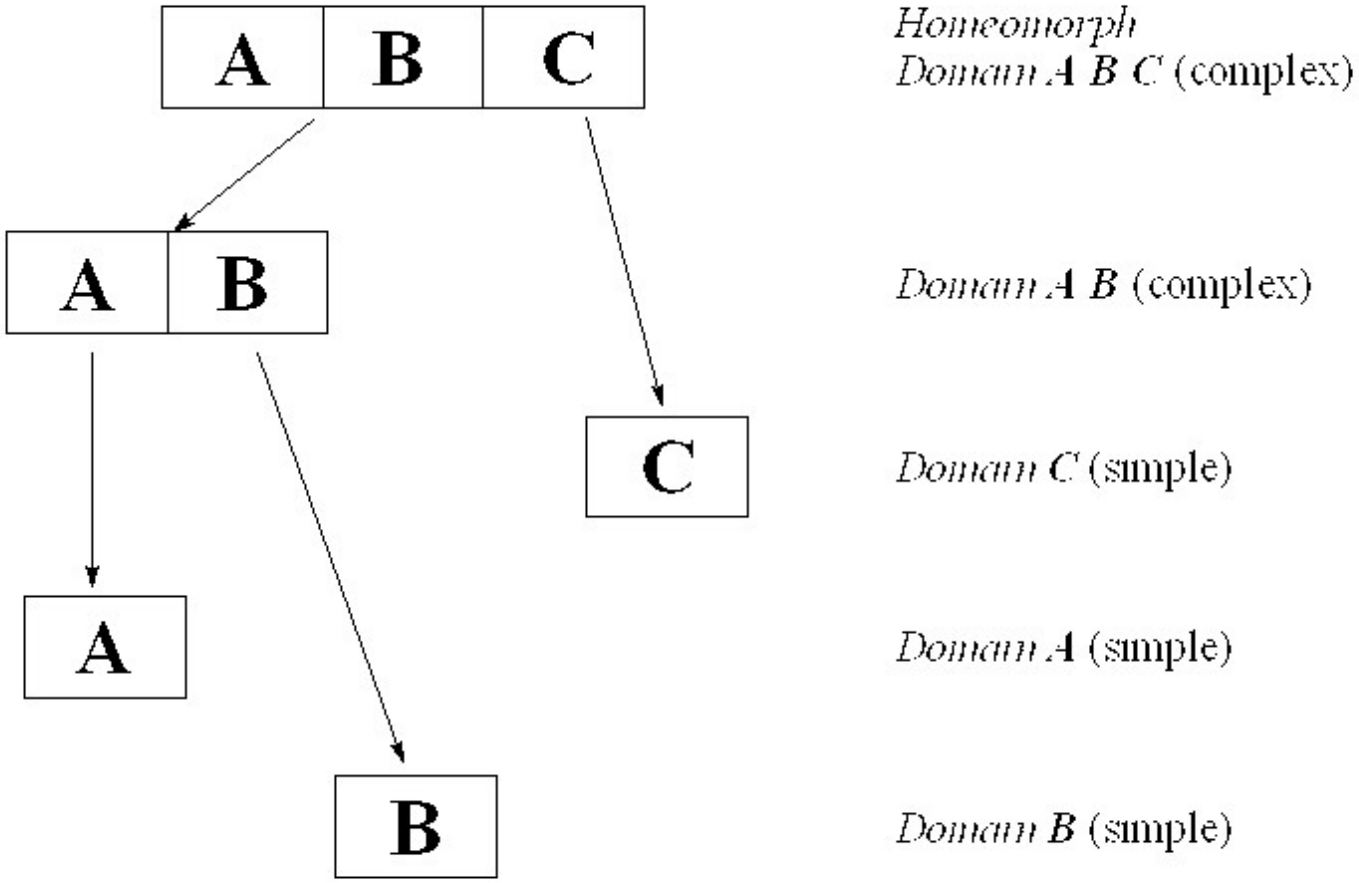
Domain classes are categorized by domain class type. The domain class type specifies a relationship(s) that characterizes the common property(s) shared by the class members.

Definition: Domain classes of type homology are called homology classes. A homology class is a class whose members have been inferred to have evolved from a common ancestor. Members of homology classes are called homologs.

Definition: A homology domain is a domain that is distinguished because it is a member of a homology class.

Definition: A sequence domain is a subsequence that has been distinguished by a well-defined set of properties or characteristics.

We are concerned here with sequence homology domains only; in the following, we will use the term domain to refer exclusively to these types of domains. Among protein sequences, it can be observed that domains often contain other domains. Consider the following example.



A protein sequence consists of domains A, B, and C. Two other domains, one containing Domains A and B (which we will call Domain A/B) the other containing Domains A/B and C (which we will call Domain A/B/C), are recognized.

To accommodate these situations we will define domains recursively, i.e., domains may contain domains that contain other domains. Domains are distinct if they correspond to different protein subsequences. For example, Domains A/B and A depicted above are distinct domains irrespective of the fact that they share a common subsequence.

Definition: A domain is said to be simple if it contains no domains as proper subsequences.

Definition: A domain is said to be complex if it contains at least one domain as a proper subsequence.

Definition: Two or more domains whose sequences are equal are said to be equal. Equal domains that are subsequences of the same sequence, that overlap in sequence position, and that are of the same domain class type are said to constitute the same domain.

The above definition requires that domains be distinguished when they are characterized as belonging to different domain class types even when they constitute the same subsequence. This distinction is necessary in order to allow unambiguous classification of sequence elements by different properties within the same classification system.

Homeomorphs

Definition: A homeomorphic class is a class of domains whose members correspond to conceptual complete sequences. Members of homeomorphic classes are called homeomorphs.

Corollary: A homeomorph is a domain that extends the entire length of the conceptual complete sequence.

Homology Families and Superfamilies

Well over 80% of the biological information concerning protein sequences reported in the published literature has been inferred by homology. Assigning sites of biological interest, features, by homology requires the construction of a multiple sequence alignment. A major goal of the PIR-International Protein Sequence Database project is to develop an objective, self-consistent system for the assignment and verification of protein sequence features by homology as depicted by multiple sequence alignments.

It has been argued by many that the construction of multiple sequence alignments is an inherently subjective process. Because mathematically rigorous multiple sequence alignment algorithms cannot guarantee biologically realistic alignments, it is common practice to adjust algorithmically generated alignments by hand. This has spawned the generation of a large number of multiple-sequence alignment editors. Nevertheless, these nonobjective alignments are used to infer important biological information that is represented in the Protein Sequence Database. We propose to employ the multiple sequence alignment as a data structure in the Protein Sequence Database. Employing such a data structure does not make the data any less objective than they already are; rather, the alignment provides a detailed record of the information that was used to make the feature assignments. Conversely, the existence of conserved biological features within the alignment provides biological verification of the correctness of the alignment.

We have observed that, among sequences and subsequences of greater than 50 residues in length that are less than 50 - 60% different, the major features of the alignment are reproduced by a wide variety of algorithms. We will refer to this as the realm of the closely related sequences. Because few gaps are required in such alignments, relatively few decisions are required for the placement of these gaps. Nevertheless, there are localized regions in such alignments, such as the immediate vicinity of the repeated glutamines in the lipoamide example given above, where the results of various methods exhibit dispersion. Fortunately, these localized differences among multiple sequence alignments are generally of little biological significance.

This realm of closely related sequences corresponds to Doolittle's sequence distance region prior to the twilight zone [8]. Significantly, within this realm, alignments derived by comparison of three-dimensional protein structures also agree well with those derived solely by sequence comparison methods [9]. We take a threshold approach to the problem of multiple sequence alignment. Closely related sequences are aligned by standard methodologies. These family groups are then further aligned using methods, such as the profile method [10], that compare the information from the entire sequence group with that from other groups in constructing the higher-order alignment. The development of superfamily alignments requires a detailed study of the biological properties of the proteins. It is our premise that fully objective and fully automatic multiple sequence alignment methods that reflect the biological understanding of the sequences can be developed at the family level.

This threshold approach is novel. Moreover, it provides a practical, objective basis for assigning and verifying protein sequence features. Within protein families there is good evidence that sequence structures are conserved. Assigning features across family groups requires an additional level of inference that cannot be justified in the absence of the additional biological information. Annotation (information related to protein sequences) is not spread automatically in the database beyond the level of protein families; descrepancies observed among annotation at the superfamily level serve as guides to focus data verification efforts.

Definition: A homology domain family is a domain family whose members are closely related under evolution. Domains are closely related if their multiple sequence alignments, as constructed by a variety of multiple sequence alignment algorithms, agree in all significant details. Members of homology families are inferred to have evolved from common evolutionary ancestors.

Homology domain families are called homology families. Note that the definition given above is deliberately vague. The goal is to provide pragmatic guidelines based on a comparison of the results of various multiple sequence alignment methods within this realm of sequence distance. Based on these results a more rigorous definition will be evolved. Definition: A homology domain superfamily is a domain superfamily composed of homology families. Members of homology superfamilies are inferred to have evolved from common evolutionary ancestors.

Definition: A homeomorphic homology family is a homology family whose members are homeomorphs.

Definition: A homeomorphic homology superfamily is a homology superfamily whose members are homeomorphs.

Members of homeomorphic homology classes are called homeomorphic homologs.

Corollary: Every protein sequence represented in the Protein Sequence Database is a homeomorphic homolog and contains at least one domain.

Proof: By the definition of subsequence, every sequence is a (improper) subsequence of itself. Provided that the sequence exhibits a well-defined characteristic, the sequence is a domain. Because the domain corresponds to a conceptual complete sequence, it is a homeomorph. Because proteins undergo continual evolution, one can postulate that every protein sequence is a homology domain. Population genetics indicates that for every protein a closely-related (possibly identical) homolog may be found somewhere within the population even if the corresponding sequence is unknown. Hence, the sequence is a homeomorphic homolog to a possibly undetermined protein sequence.

Corollary: Sequence homology is a relationship that allows all protein domains to be partitioned and each partition is closed under transitivity. Hence, homology families and homology superfamilies can be defined.

Proof: Because every domain is homologous to itself (a sequence domain in the database represents a class of sequences that occur in an entire population, each member of which is homologous to the other corresponding members), every domain can be assigned to at least one homology class. We will restrict the assignment of domains to domain classes, such that a domain may be assigned to a single class only, that to which it is most closely related. Hence, the classification will constitute a partitioning of all domains represented in the Protein Sequence Database. Because all members of the class are mutually homologous, transitive closure is ensured. Because every domain can be assigned to a homology class once it has been elucidated, it can be inferred that all domains can be classified accordingly.

In practice, a concerted effort will be made to assign sequences in the Protein Sequence Database to homeomorphic homology families and superfamilies only when at least two members of the class can be identified or when there is some other specific reason to do so. By default, sequences that have not been assigned will tentatively be understood to constitute homeomorphic classes unto themselves. Domains that are proper subsequences of the sequences represented in the database will be assigned to homology classes as they are identified. Note that because domains are considered to be distinct even when they contain overlapping segments, overlapping domains (from the same protein) must be assigned to different homology classes.

Effects on the Database Model

The information in the PIR-International Protein Sequence Database is being separated into a set of interlinked components. The essential strategy behind the component model for protein classes is that these groupings can be compiled separately from the sequence entries in the database. The information in the Class Component will be used 1) to automatically generate the superfamily line in the sequence entries, 2) to order the sequence entries within homeomorphic classes, 3) to provide a platform for comparing and reconciling annotations (features included) among the members of each class, and 4) to transfer (spread) these refined annotations back to the entries in the sequence component.

The strategy for compiling these classes is based on the premise that clusters of domains will be compiled at the family level and that superfamilies will be assembled from these families. In practice, it is often easier to recognize that the sequences or subsequences are members of the same superfamily than to partition them into families and to align the families. Hence, the data structure adopted for representation of protein classes allows for nonaligned members and does not require that all members be assigned to families. Note that this is consistent with the definition of homology superfamily and family because any domain can be interpreted as a family containing a single member.

Superfamily and family classes share a common representational structure. All protein sequence entries in the database whose sequences have been assigned to homeomorphic superfamilies will be ordered by homeomorphic class (unassigned sequence entries will follow ordered by species); this provides a natural order to the database. Hence, it is necessary to specifically designate entries in the Class Component that correspond to homeomorphic classes.

A Class Component entry models that of a Protein Sequence Component entry. All records defined for a Protein Sequence Component entry may be included in a Class Component entry when appropriate and nearly all are optional (the rules for appropriateness and nearly all have not been elucidated). There are four important differences from the Protein Sequence Component entries: (1) records in the class entries are hierarchically nested; (2) the class entries may include **ALIGN** and **MEMBERS** records that specify the member domains; (3) the hierarchical structure of the records associates characteristics and properties depicted on the class records with specific sets of member domains; (4) feature records refer specifically to aligned class members as depicted on the ALIGN records.

The following is an example of a class entry depicting the lactate/malate dehydrogenase homeomorphic superfamily.

```
{{ENTRY          GP1037#class homeomorph
                      #level superfamily #domain-type simple }
 {TITLEl         actate/malate dehydrogenase }
 {KEYWORDS       NAD; oxidoreductase; tetramer}
 {BINDING_SITE   107 #description coenzyme #ligand NAD }
 {
   {SUBCLASS     #description L-lactate dehydrogenase (EC 1.1.1.27),
                      animal #level family }
   {BINDING_SITE 100,170 #description substrate #ligand lactate }
   {
     { SUBCLASS  #description H-chain }
     { ALIGN     DEPGLH #gaps 0/1 #checksum 6356 \
                 DECHLH #gaps 0/1; 18/1 #checksum 1403 }
     }
   {
     { SUBCLASS  #description M-chain }
     { ALIGN     DEHULM #gaps 18/1; 333/1 #checksum 7516 }
     { MEMBERS   DEMSLM }
     { ALIGN     DEPGLM  #gaps 0/1; 18/1; 333/1 #checksum 2715 }
```

```
      { MEMBERS    DECHLM }
      }
    ...
    }
  ...
}
```

ALIGN records give a sequence specification, a list of the gaps in the sequence, and a checksum on the sequence defined by the sequence specification. The sequence specification may include a database component identification code (PIR1 for example), an entry identification code, a record identification code, and a segment specification. The segment specification specifies a subsequence of the sequence indicated by entry identification code. The record identification code identifies a record in the sequence database entry that contains a segment specification. The database identification code, the record identification code, and the segment specification are optional; the entry identification code is required. For example, PIR1:DEHULM->MAT, DEHULM(35-107), and DEHULM are valid sequence specifications. If a segment specification is used on an ALIGN (or MEMBERS record, see below) corresponding to a homeomorph, it must specify a subsequence that is identical with the sequence shown in the sequence database entry. ?????22 The segment specification, in general, allows a sequence to be transformed using the full set of insertion, deletion, substitution, duplication, and transposition operations.?????

MEMBERS records give a list of sequence specifications that correspond to members of the class. Domains specified on ALIGN records are also members of the class. MEMBERS records may be used to register domains as members of the class without explicit alignment of the domains. Domains as described by the sequence specification on MEMBERS records are assumed to be incompletely defined even when a segment specification is present or indicated.

In other words, a sequence depicted on a MEMBERS record is understood to contain a domain, whose location within the sequence or whose boundaries have not been determined with confidence; it is that domain that is a member of the class. Even when the location or boundaries are specified explicitly they are understood to be approximate. Note that because one cannot confirm the boundaries of a homology domain without aligning it with other domains, this interpretation places no restrictions on the data. Homeomorphic domains are an exception: they are known and expected to comprise the entire sequence.

The entry class (ENTRY #class) may contain either the value homeomorph or the value domain only; when unspecified it is assumed to be domain. The class level (ENTRY #level) may contain either the value superfamily or the value family only; the default is unspecified, i.e., the class represents an unspecified sequence grouping. The domain type (ENTRY #domain-type) may contain the value complex or simple only; the default is simple. Each level in the hierarchy (with the exception of the root) is introduced by a SUBCLASS record. Optionally, SUBCLASS records may contain a #level field; when present this field may contain the value family only. Although this structure allows classes to be nested to as many levels as desired, only the superfamily and family levels are distinguished. A class-type field (ENTRY #class-type) is also defined. Currently, the only valid value is homology; by default this value is assumed. This field has been included to allow for the future inclusion of other class-type entries in the Class Component.

The inclusion of the ENTRY #class field permits all homeomorphic superfamilies and families to be selected; the inclusion of the ENTRY #domain-type field permits the set of all simple domain superfamilies and families to be selected (which permits nonredundant sequence sets to be assembled).

Member domains are ordered within superfamilies and families in the order in which they are depicted on ALIGN or MEMBERS records. In the above example, the order is DEPGLH, DECHLH, DEHULM, DEMSLM, DEPGLM, and DECHLM.

The database may be ordered by ordering the homeomorphic superfamily classes (a separate component will be needed to store this order). An order will also be established for complex superfamily classes and for simple superfamily classes. These orderings will allow the Superfamily Records in the Protein Sequence Component entries to be generated automatically from the Class Component entries in a well-established order. Superfamily Records contain a list of the names of all superfamilies to which domains within the sequence represented in the entry belong. These will be listed in the order, homeomorphic superfamily first followed by complex domain superfamilies and then by simple domain superfamilies. Note that as long as a homeomorphic superfamily is constructed whenever a domain from the sequence is included in a nonhomeomorphic superfamily it can be guaranteed that the first and only the first superfamily name on the Superfamily Record corresponds to the homeomorphic superfamily.

MEMBERS records may also contain specifications for other entries in the Class Component (technically these specifications fall under the class of sequence specifications although they do not specify a unique sequence) rather than for sequences in the Sequence Component. This formalism may be used to include family entries into superfamily entries only. This facility allows family alignments to be developed individually; members of these alignments can be incorporated into superfamilies directly. Note that the linkage between superfamilies and families through the MEMBERS record was designed specifically to avoid the complications involved in aligning alignments of sequences, initially. This capability may be added at a later time.

Mathematical Definition of Homeomorphism

The following has been included as an explanation of the choice of the term homeomorph to mean a member of a class of conceptual complete sequences.

A function is a mapping from domain X to range Y such that if x is in the space X then f(x) is in the space Y. The function f is said to be continuous at x if, for every epsilon > 0, there is a delta > 0 such that if | x - z | < epsilon then | f(x) - f(z) | < delta for all z in domain X. The function f is called continuous if it is continuous at each x in its domain X. A function f from domain X to range Y is one-to-one if and only if f(x) = f(z) when x = z. A mapping of X is onto Y if for all y in Y there is some x in X such that f(x) = y. Functions that are one-to-one and onto are often called one-to-one correspondences between X and Y. In these cases, there is a function g mapping Y to X such that, for all x and y, g(f(x)) = x when f(g(y)) = y. This function is called the inverse function of f [11].

A one-to-one mapping f of X onto Y is called a homeomorphism between X and Y if f is continuous and the inverse function f is also continuous. The spaces X and Y are said to be homeomorphic if there is a homeomorphism between them [11].

Although a homeomorphism is defined as a relationship between spaces in general rather than ordered sets or sequences of discrete elements, it captures the essential idea of interest here; it describes a complete cross-mapping of all elements of two or more sequences. In this case, the homeomorphism is the function that transforms the residues of sequence A to sequence B by a series of insertion, deletion, or substitution operations: the listing as defined by Kruskal [6]. A sequence alignment specifies a unique listing and vice-versa, the two representations are equivalent except that an alignment is symmetric, whereas a listing is asymmetric [6].

References:

1. Dayhoff, M.O., Computer analysis of protein sequences, Fed. Proc. 33, 2314-2316, 1974.
2. Dayhoff, M.O., McLaughlin, P.J., Barker, W.C., and Hunt, L.T., Evolution of sequences within protein superfamilies, Naturwissenschaften 62, 154-161, 1975.
3. Dayhoff, M.O., The origin and evolution of protein superfamilies, Fed. Proc. 35, 2132-2138, 1976.
4. Webster's New Collegiate Dictionary, 1975, G. & C. Merriam Co.
5. Batini, C., Ceri, S., Navathe, S.B., Conceptual database design, an entity-relationship approach, The Benjamin/Cummings Publishing Company, Inc., Redwood City, CA, 1992.
6. Kruskal, J.B., An overview of sequence comparison, In: Time warps, string edits, and macromolecules: the theory and practice of sequence comparison, Sankoff, D., and Kruskal, J.B., eds., pp. 1-44, Addison-Wesley Publishing Company, Inc., 1983.
7. Kruskal, J.B. and Liberman, M., The symmetric time-warping problem: from continuous to discrete, In: Time warps, string edits, and macromolecules: the theory and practice of sequence comparison, Sankoff, D., and Kruskal, J.B., eds., pp. 125-161, Addison-Wesley Publishing Company, Inc., 1983.
8. Doolittle, R.F., Of URFs and ORFs: a primer on how to analyze derived amino acid sequences, University Science Books, Mill Valley, CA, 1987.
9. Sander, C., and Schneider, R., Database of homology-derived protein structures and the structural meaning of sequence alignment, PROTEINS 9, 56-68, 1991.
10. Gribskov, M., McLachlan, A.D., Eisenberg, D., Profile analysis: detection of distantly related proteins, Proc. Natl. Acad. USA 84, 4355-4358, 1987.
11. Roydon, H.L., Real Analysis, Second Edition, MacMillan Publishing Co., Inc., New York, 1968.