The database is made available directly from the PIR-International Centers and from a variety of sites on the World Wide Web. For specific information contact the PIR Technical Services Coordinator, National Biomedical Research Foundation, 3900 Reservoir Road NW, Washington, DC 20007; telephone +1 202 687-2121; FAX +1 202 687-1662; electronic mail PIRMAIL@nbrf,georgetown.edu. In Europe, contact MIPS: Martinsried Institut für Proteinsequenzen, Max-Planck-Institut für Biochemie, D-82152 Martinsried bei München, Germany; telephone +49 89 8578 2657; FAX +49 89 8578 2655; electronic mail MIPS@ehpmic.mips.biochem.mpg.de. In Asia or Australia, contact JIPID: Japan International Protein Information Database, Science University of Tokyo, 2669 Yamazaki, Noda 278, Japan; telephone +81 471 239778; FAX +81 471 221544; electronic mail TSUGITA@JPNSUT31.bitnet or EX5292@JPNSUT30.bitnet.

Acknowledgments

We acknowledge the contributions of all of the scientists and staff of the PIR-International centers whose work we report here. This publication was supported in part by Grant P41 LM05798 from the National Library of Medicine. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Library of Medicine.

[4] Superfamily Classification in PIR-International Protein Sequence Database

By Winona C. Barker, Friedhelm Pfeiffer, and David G. George

Brief History of Superfamily Concept

In the mid 1970s, Dayhoff proposed that all naturally occurring proteins would cluster into families and superfamilies whose members have diverged from common ancestral forms. A similar proposal was made by Emil Zuckerkandl. Estimates of the number of protein superfamilies were in the low thousands. Using a variety of criteria for superfamily membership,

more recent estimates of the same order of magnitude have been reported.\textsuperscript{4-6}

Although superfamily relationships sometimes are so ancient as to preclude recognition solely on the basis of sequence similarity, our group has employed sequence similarity as the main criterion for partitioning the Protein Sequence Database into independent, nonoverlapping groups. In 1976, the nearly 500 completely sequenced proteins then known were each assigned to one of 116 superfamilies.\textsuperscript{7} At that time, there were no examples in the database of complete precursor sequences, of polyproteins, or of products of alternative splicing of mRNA. Most of the known sequences were of mature forms of peptides or proteins. There were only a few examples of multidomain proteins.

Within a few years of the introduction of the superfamily concept, it became evident that many protein sequences contain homology domains, namely, regions of local similarity contained in otherwise unrelated proteins. Such domains are often responsible for similar properties (such as calcium binding, DNA binding, or catalytic activity) shared by diverse proteins. Evidence from X-ray crystallography and chemical studies revealed that these domains often correspond with compact regions of the structure or with easily cleaved fragments. More surprising was the discovery that the genes for many proteins contain noncoding regions (introns) that divide the protein coding region into exons that sometimes approximately correspond with the domains as defined by structure or protein sequence. "Exon shuffling" among genes is now recognized as one mechanism in the evolution of "new" proteins.

Although the term protein superfamily is widely used, its meaning is not well defined when applied to multidomain proteins. In the literature, terms such as "the immunoglobulin superfamily"\textsuperscript{8,9} have come to mean the collection of all proteins that contain the named domain. This usage of the term, however, does not allow the database to be unambiguously partitioned into superfamilies. Multidomain proteins such as the platelet-derived growth factor receptor can be placed in the protein kinase superfamily, or in the immunoglobulin superfamily, or in a different superfamily of se-

quences that contain both of these types of domains. We have developed a formal model of the protein superfamily concept that encompasses the most common usages and integrates both homology at the domain level and homology at the level of complete proteins.\textsuperscript{10,11} This model preserves the ability to fully partition the Protein Sequence Database, permits the organization of the database in a structured way, and introduces a more precise and unambiguous definition for the term protein superfamily.

Superfamily Concept Revised and Generalized

The concepts of superfamily and family have been generalized to encompass any scheme for classifying proteins (or regions within proteins) that partitions the proteins (or protein regions) into hierarchically nested sets that are closed under transitivity.\textsuperscript{10,11} A superfamily is a union over families. Families are sets within the superfamily hierarchy for which the members meet a threshold level of relatedness.

In the Protein Information Resource (PIR)-International Protein Sequence Database we classify sequence homology domains and apply the terms superfamily and family to these units of information. A homology domain is a sequence region found in diverse proteins that is likely to be derived from a common evolutionary ancestor. Homology domains differ from patterns or motifs (that may be contained within them) in that they are demonstrably similar along their entire extents as observed by multiple sequence comparison and alignment. Generally, they are greater than 50 residues in length. The most common homology domains noted in the June 1995 release of the Protein Sequence Database are listed in Table I.

A homology domain may encompass all of a protein sequence or may represent a subsequence within it. Because the insertion (or deletion) of exogenous loops of sequence within individual domains is a commonly observed evolutionary event, the subsequence composing a homology domain may not be contiguous. Protein domains may be complex, that is, composed of more than one distinct domain. Complex domains may be formed by coalescence of two or more originally independently evolving domains; after concatenation, the domains evolve as a unit. Domains not composed from other identifiable domains are called simple domains. Examples of some complex domains are listed in Table II. Protein sequences may contain a single simple domain or may be mosaics composed of a

\textsuperscript{10} D. G. George, 1993, unpublished.

### TABLE I
**Most Commonly Occurring Homology Domains**

<table>
<thead>
<tr>
<th>Homology domain</th>
<th>Number of entries&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Taxonomic distribution&lt;sup&gt;c&lt;/sup&gt; (%)</th>
<th>Animals&lt;sup&gt;d&lt;/sup&gt;</th>
<th>Plants&lt;sup&gt;e&lt;/sup&gt;</th>
<th>Fungi</th>
<th>Protists</th>
<th>Prokaryotes&lt;sup&gt;f&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Immunoglobulin</td>
<td>2922</td>
<td></td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Protein kinase</td>
<td>623</td>
<td></td>
<td>85</td>
<td>5</td>
<td>7</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Homeobox</td>
<td>531</td>
<td></td>
<td>96</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Calmodulin repeat</td>
<td>428</td>
<td></td>
<td>86</td>
<td>6</td>
<td>3</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Trypsin</td>
<td>339</td>
<td></td>
<td>98</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Kazal proteinase inhibitor</td>
<td>191</td>
<td></td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Myosin head</td>
<td>176</td>
<td></td>
<td>88</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Cytochrome c</td>
<td>166</td>
<td></td>
<td>32</td>
<td>23</td>
<td>9</td>
<td>3</td>
<td>33</td>
</tr>
<tr>
<td>SH3 (src homology 3)</td>
<td>158</td>
<td></td>
<td>89</td>
<td>0</td>
<td>8</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>Ribonucleoprotein repeat</td>
<td>149</td>
<td></td>
<td>72</td>
<td>15</td>
<td>12</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>H&lt;sup&gt;+&lt;/sup&gt;-transporting ATP synthase α chain</td>
<td>146</td>
<td></td>
<td>16</td>
<td>41</td>
<td>10</td>
<td>1</td>
<td>32</td>
</tr>
<tr>
<td>Translation elongation factor Tu</td>
<td>145</td>
<td></td>
<td>16</td>
<td>19</td>
<td>12</td>
<td>5</td>
<td>48</td>
</tr>
<tr>
<td>EGF (epidermal growth factor)</td>
<td>134</td>
<td></td>
<td>99</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MalK protein</td>
<td>131</td>
<td></td>
<td>34</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>55</td>
</tr>
<tr>
<td>SH2 (src homology 2)</td>
<td>129</td>
<td></td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<sup>a</sup> In the database, the names of homology domains end with the word "homology," which has been omitted in these tables for brevity.

<sup>b</sup> Number of entries containing domain found in PIR-International Protein Sequence Database, Release 45.0 (June 1995), with sections PIR1 and PIR2 totaling 66,573 entries.

<sup>c</sup> Percentage of entries found (column 2).

<sup>d</sup> Includes animal viruses.

<sup>e</sup> Includes plant viruses.

<sup>f</sup> Includes bacteriophages.

The domain that represents the entire protein is called the homeomorphic domain. Because the precursor form (the form initially translated from the mRNA) of the protein is generally represented in the database, homeomorphic domains include portions of the sequence removed during maturation. Two proteins belong to the same homeomorphic superfamily when they show homology over the length of their entire sequences; hence, two members of the same homeomorphic superfamily contain the same homology domains in the same order.

Within a homology domain superfamily, more closely related domains...
<table>
<thead>
<tr>
<th>Complex domain</th>
<th>Constituent domains</th>
<th>Proteins&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cytochrome b</td>
<td>Cytochrome b&lt;sub&gt;6&lt;/sub&gt;</td>
<td>FcbH bifunctional protein [B32382]</td>
</tr>
<tr>
<td>Plastoquinol–plastocyanin reductase 17K protein</td>
<td></td>
<td>Cytochrome b [CBHU]</td>
</tr>
<tr>
<td>Carbamoyl-phosphate synthase (ammonia)</td>
<td>Carbamoyl-phosphate synthase (glutamine-hydrolyzing) large chain</td>
<td>Carbamoyl-phosphate synthase (ammonia) I [SYRTCA]</td>
</tr>
<tr>
<td></td>
<td>Carbamoyl-phosphate synthase (glutamine-hydrolyzing) small chain</td>
<td>Pyrimidine synthesis protein CAD [A23443]</td>
</tr>
<tr>
<td></td>
<td>TrpG</td>
<td></td>
</tr>
<tr>
<td>HisI bifunctional enzyme</td>
<td>HisI protein</td>
<td>HisI–hisD trifunctional enzyme [SHNC]</td>
</tr>
<tr>
<td></td>
<td>Histidinol dehydrogenase</td>
<td>HisI bifunctional enzyme [YNECHI]</td>
</tr>
<tr>
<td>Leukocyte common antigen cytosolic domain</td>
<td>Protein-tyrosine-phosphatase (2 copies)</td>
<td>Leukocyte antigen-related protein [TDHULK]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Leukocyte common antigen [A46546]</td>
</tr>
<tr>
<td>Osteonectin homology</td>
<td>Agrin inhibitor-like repeat</td>
<td>Matrix glycoprotein SC1 [GERTX1]</td>
</tr>
<tr>
<td></td>
<td>Kazal proteinase inhibitor</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Calmodulin repeat</td>
<td>Osteonectin [GEHUN]</td>
</tr>
</tbody>
</table>

<sup>a</sup> Entry code or accession number is given in square brackets.

are grouped into families. For practical reasons we place domains into the same protein family if they show at least 50% sequence identity (see [3], this volume). When the data are fragmentary, they may be classified provided that one or more homologs exist that are related closely enough to reasonably assume that the missing data conform with those of the homologs.

Partitioning of the families and superfamilies is achieved by treating homology domains containing overlapping regions independently; in other words, complex domains and the simple domains from which they are composed are separately classified and treated as independent entities. This redundancy allows the entire database to be partitioned into homeomorphic families while simultaneously allowing the entire (nonredundant) collection of simple domains to be independently classified. A system is under development for maintaining the relationships among simple and complex domains separate from the superfamily classification scheme.
### TABLE III
PROTEINS WITH FIVE OR MORE TYPES OF HOMOLOGY DOMAINS

<table>
<thead>
<tr>
<th>PIR entry</th>
<th>Superfamily name</th>
<th>Simple homology domains</th>
</tr>
</thead>
</table>
| XYRTFA, fatty-acid synthase (EC 2.3.1.85)—rat | Rat fatty-acid synthase | 3-Oxoacyl-[acyl-carrier-protein] synthase I  
Acyl carrier protein  
Long-chain alcohol dehydrogenase  
Oleoyl-[acyl-carrier-protein] hydrolase  
Short-chain alcohol dehydrogenase  
[Acyl-carrier-protein] S-malonyltransferase |
| BVBYA1, ARO1 protein—yeast (Saccharomyces cerevisiae) | Aro1 protein | 3-Dehydroquinate dehydratase  
3-Dehydroquinate synthase  
3-Phosphoshikimate 1-carboxyvinyltransferase  
Shikimate dehydrogenase  
Shikimate kinase |
| BVASA1 aroM protein—Emericella nidulans |  |  |
| KFHUI2, coagulation factor XIIa (EC 3.4.21.38) precursor—human | Coagulation factor XII | EGF (epidermal growth factor)  
Fibronectin type I repeat  
Fibronectin type II repeat  
Kringles  
Trypsin |
| QZFF, rudimentary protein—fruit fly (Drosophila melanogaster) | Rudimentary enzyme | Aspartate/ornithine carbamoyltransferase  
Bacillus dihydroorotase  
Carbamoyl-phosphate synthase (ammonia)  
Carbamoyl-phosphate synthase (glutamine-hydrolyzing) large chain  
Carbamoyl-phosphate synthase (glutamine-hydrolyzing) small chain |
| A23443, pyrimidine synthesis protein CAD—golden hamster |  |  |
| QZBYU2, pyrimidine synthesis protein URA2—yeast (Saccharomyces cerevisiae) |  |  |
| QZDOP3, pyrimidine synthesis protein PYR1-3—slime mold (Dictyostelium discoideum) (fragments) |  |  |

### Defining Homeomorphic Families and Superfamilies

For purposes of partitioning the databases, the defining relationship is said to be sequence homology, that is, the logical inference of common ancestry. However, homology is not a quantifiable characteristic: there either was common ancestry or there was not. There are no degrees of homology; thus, no thresholds can be established reliably to detect homology or to partition sequences within homology groups. In practice, we use...
similarity as an indicator of homology and establish threshold levels of similarity on which to base the partitioning.

Methods of detecting and quantifying sequence similarity range from fairly straightforward to rather sophisticated and computationally expensive. The latter methods, however, are needed only for the difficult cases—those where the sequence similarity is not well conserved. Although these very distant relationships are scientifically of great interest, for the purposes of database operations (classification, standardization of annotation within classes) such clusterings are less useful than clusterings of closely related sequences. There is much empirical evidence suggesting that a threshold can be established whereby it can be inferred that closely related proteins share common biological properties. Because closely related groups of proteins, which generally are homologs of the same protein in various species or products of recent gene duplications, can be expected to share many structural and functional characteristics, new sequences can directly inherit annotation information associated with existing closely related sequences. This provides a mechanism for comprehensive and consistent annotation within large areas of the database.

Formerly, our approach to classification was to group sequences into superfamilies, which were then subdivided into families, subfamilies, entries, and subentries according to percentage of sequence identity. This has the disadvantage that it may involve making difficult (and time-consuming) decisions prior to making simple and very useful decisions that are amenable to automation.

We now initially classify proteins into homeomorphic protein families. To facilitate the development of semiautomated procedures, we have adopted a working definition of a homeomorphic protein family as a set of sequences that can be aligned end to end without major discrepancy by standard multiple sequence alignment methods. In practice, such sequences generally will have an overall sequence identity of at least 50% and the same domain architecture. Classification into protein families is one of the first steps in annotation.

Protein families for which the members have similar overall architecture (i.e., the same domains in the same order) are then clustered into homeomorphic superfamilies. Sequences representing different families within a superfamily are alignable end to end but with less certainty and with considerably more latitude (e.g., we permit terminal regions and interdomain regions that are sufficiently dissimilar that the alignment is essentially arbitrary, some variability in the number of repeats of a domain and in the size of regions of restricted composition, and the absence of a domain because of alternative splicing or use of an alternative initiator).
Clustering Sequences into Protein Families

The FASTA program permits rapid comparison of a query sequence to a sequence database. The Martinsried Institute for Protein Sequences (MIPS), the European branch of PIR-International, has generated the FASTA database. This database contains results of FASTA searches of each entry in the database against the entire PIR-International Protein Sequence Database (plus the PATCHX database of sequences available elsewhere but not yet processed by PIR-International). The FASTA database is updated dynamically as new entries are added to the database.

The FASTA results can be used either to select candidates for an existing protein family from among as yet unclassified entries or to search for the protein family to which a single unclassified entry belongs. Candidate sequences are aligned automatically, and the aligned pairs are screened for congruence of length and threshold level of similarity. Those that meet rather stringent requirements are routinely classified; others are examined and classified by scientific staff. Multiple sequence alignments are computed for all protein families containing more than two members, using the Feng and Doolittle algorithm as implemented in the Genetics Computer Group PILEUP program.

As of release 45.0 of the Protein Sequence Database (June 1995), 70% of all database entries have been classified into protein families. About 7% of all sequences are not classifiable, generally because the sequences are too short or are fragmentary. About 18% of all classified entries are unique representatives for a protein family. Routinely, when a protein family contains members that have been assigned to a superfamily, all members of the family are automatically classified into that superfamily.

Placement Group Classification

Placement is defined as the ordering of the sequences (homeomorphic domains) in the database. We strive to place homeomorphic superfamilies into broader categories: electron-transfer proteins, enzymes, enzyme inhibitors, etc. Within these categories, we often cluster superfamilies that contain one or more related domains, have some functional relationship (e.g., participation in the same physiological process), or possess some other biological similarity.

A series of five numbers is used to specify the placement of each sequence entry. The first number identifies the placement (homeomorphic superfamily) group, the second identifies the family within the superfamily group, and the subsequent numbers identify subfamily, entry, and subentry. The subfamily and entry numbers were originally employed as an indication of threshold levels of sequence similarity (more than 80% and 95% identical, respectively). Currently, they are used as a convenience in ordering large superfamilies, without strict adherence to the numerical thresholds. The number 0.0 is assigned at the family level and below to indicate a null assignment when a strict ordering within the superfamily or family has not been established. Our goal is to classify all classifiable database entries at the family and superfamily levels. However, classification will be tentative for entries containing fragmentary data or for peptides derived from much larger precursors. In each database release, unclassified sequences follow the classified sequences; they are sorted first by taxonomic class and then ordered alphabetically by entry title.

The placement numbers are not stable identifiers for the superfamilies. When a new superfamily or subgroup within a superfamily is required, we may use a fractional number. From time to time, the database is renumbered and all numbers are converted to consecutive whole numbers. Therefore, it is recommended that the superfamily name (see below) be used when referring to a superfamily.

Defining Homology Domains

The essential concept of homology domains predated the development of the superfamily concept; it arose from the observation that duplicated regions within homologous proteins showed more similarity to the corresponding regions of their homologs than to the duplicated region within the same protein sequence. This evidence of gene duplication was first described for ferredoxins. In the Protein Sequence Database, the term “homology regions” was first applied in the study of the evolution of the C (constant) regions of the classic immunoglobulins. In the mid 1970s the National Biomedical Research Foundation (NBRF) introduced a specialized program, RELATE, for detecting sequence duplication. All modern database searching programs are designed to detect “local” sequence similarities and are well suited for detecting unusually similar regions or domains.

in otherwise unrelated sequences. Homology domains are now commonly observed by researchers examining the results of database searches.

Elucidating the relationships among domains is a more difficult subset of the task of discovering distant protein relationships. There is a large literature on methods for examining such relationships, and these methods continue to be studied and improved (see, e.g., Pearson\textsuperscript{18}). A full analysis of the significance of proposed domain relationships can require the use of a number of the most sophisticated tools available for computer analysis of sequences. Confirmation of proposed structural and functional relationships may require nuclear magnetic resonance (NMR) or x-ray crystallographic studies, site-directed mutagenesis, and analysis of the properties of chemically modified or engineered sequences. Therefore, the full characterization of any type of protein domain is beyond the scope of the activities of a sequence database. All of the homology domains annotated in the Protein Sequence Database are understood to have been deduced on the basis of sequence similarity, and only sometimes confirmed by other evidence. When we "define" a homology domain, we are not claiming to have characterized it; we are simply recording the criteria that we use to determine that such a domain does exist in the protein.

From the beginning, we have defined a homology domain by constructing a multiple sequence alignment of the proposed homologous segments. When refining and evaluating such an alignment, we also consider other information, such as the identity and location of known functional residues. Because homology domains of a given type can be very distantly related (less than 20% identical in some cases), the details of such alignments, outside of regions that are well conserved, are highly tentative. We periodically review our domain alignments in the light of new examples that have been discovered and of structural and functional relationships that have been elucidated by experimental research.

When a pattern of conserved residues or sequence characteristics is clear, the major decision to be made is the assignment of the boundaries of the domain. Often the similarity is weak at one or both ends, and the length of intradomain regions in proteins where the domain is repeated is variable. Thus, the boundary assignments may be rather arbitrary and differ from author to author. We take the view that domain boundaries are approximate and to some extent arbitrary. For users of the database, the presence of a domain is more useful information than are estimates of its "real" boundaries. Rather than trying to portray the maximum extent of a domain, which may involve aligning regions of little similarity, we often choose boundaries in relation to conserved features that occur close to the

\textsuperscript{18} W. R. Pearson, Protein Sci. 4, 1145 (1995).
ends of the domain. This strategy has several advantages: the choice is more often obvious, and little time need be spent making boundary decisions; homologous domains can be reasonably aligned using widely available multiple alignment software; and computer-assisted methods for detecting additional examples of the known homology domains can be used effectively. For the maintenance of a database of rapidly increasing size, these practical considerations are very important.

We attempt to annotate self-consistently all occurrences of a homology domain in the feature records using a semiautomated procedure. A computer program examines the PIR-International Protein Sequence Database and extracts each annotated homology domain. With the exception of signal sequences, transit peptides, short sequences, and short interdomain regions (<50 residues), all sequences and subsequences not represented in this data set are also extracted. These segments and the identified homology domains are cross-compared and the intercomparison scores are stored in a FASTA database. These data are examined to ensure consistency in domain assignments and to detect previously unidentified members of existing domain families. Segments that show high overall sequence similarity with a known homology domain are automatically aligned with the most similar member of the domain family. If the alignment shows high conservation at both ends, does not contain large gaps, and shows a uniform distribution of sequence similarity, then the segment is automatically classified as an additional member of the domain family. These criteria aid in preserving the homology domain boundaries as originally established. A multiple alignment of the entire family is then computed and the domain boundaries are refined accordingly. Additional segments overlooked by this procedure may be included in the family after examination of the FASTA scores by database staff.

Naming of Superfamilies and Domains

Names are assigned to each homeomorphic superfamily and to each homology domain. These names are given in the Protein Sequence Database on the Superfamily record. If the sequence has been classified into a homeomorphic superfamily, its name appears first on the record. The names of homology domains are distinguished from homeomorphic superfamily names by including the word "homology" at the end of the name.

Usually the superfamily or domain is initially named for the first or most well-known member protein characterized. We often must assign a name when function is poorly characterized and it is not known how wide-

spread is the occurrence of the protein or domain. We strive to make the names recognizable, either because they are descriptive (hevein chitin-binding domain homology) or because they are commonly used (kringle homology). A domain that is repeated in the protein for which it is named is labeled a repeat homology. Sometimes the name of a superfamily or of a homology domain is later changed to a more suitable name or to one that has become widely used.

Conclusions

Since the origin of the Protein Sequence Database, the classification of protein sequences into superfamilies has provided a biologically meaningful organization of the data. As discussed in [3] in this volume, this classification provides a systematic scheme for verification of the information in the database and for inferring additional information by homology in a controlled way. Information generated from large-scale sequencing projects is incomplete and not well understood. The major task of computational biology is to assign biological meaning to these data. Homology is the major operating principle employed in these analyses. The superfamily classification provides a useful architecture for self-consistent and objective examination of sequence data by homology.

Availability of Data

The classifications as presented in this chapter are available from the nodes of PIR-International in various ways. The Superfamily record and homology domains annotated as features are integral parts of the Protein Sequence Database entries. Retrieval programs available from PIR-International, ATLAS and PSQ, allow selection of sequence entries by superfamily name or placement number. On-line access is possible both at NBRF and at MIPS. These sites also operate full-function network file servers that handle database queries, sequence searches, and sequence submissions, in addition to file server requests.

For further information, please contact the PIR Technical Services Coordinator, National Biomedical Research Foundation, 3900 Reservoir Road NW, Washington, DC 20007; telephone +1 202 687-2121; FAX +1 202 687-1662; electronic mail PIRMAIL@nbrf.georgetown.edu. In Europe, contact MIPS: Martinsried Institute for Protein Sequences, Max Planck Institute for Biochemistry, D-82152 Martinsried near Munich, Germany; telephone

The Protein Sequence Database is also accessible on the Internet over the World Wide Web (WWW) (web address for MIPS: http://www.mips.biochem.mpg.de/). The results from classification into protein families are also available via the WWW at MIPS. As of June 1995, about 6000 multiple alignments including 40,000 database entries were available. The multiple alignments are enriched by addition of the consensus sequence and of all features annotated in the database entries. In addition, 230 multiple alignments of homology domains, representing about 10,000 individual homology domains, can be inspected. Additional WWW services will be implemented in the near future; contact the MIPS WWW site or any of the PIR-International nodes for further information.

Acknowledgments

We gratefully acknowledge the contributions of all of the scientists and staff of the PIR-International centers whose work we report here. This work was supported in part by Grant P41 LM05798 from the National Library of Medicine. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Library of Medicine. MIPS is supported by the Max-Planck-Gesellschaft, the Forschungszentrum f. Umwelt und Gesundheit (GSF), and the European Economic Community BRIDGE Programme Grants BIOT-CT-0167 and BIOT-CT-0172.

[5] Gene Classification Artificial Neural System

By Cathy H. Wu

Introduction

As technology improves and molecular sequencing data accumulate exponentially, continued progress in the Human Genome Project will depend increasingly on the development of advanced computational tools for rapid annotation and easy organization of genomic sequences. Currently, a database search for sequence similarities represents the most direct computational approach to decipher the codes connecting molecular sequences with protein structure and function. A sequence classification method can