minireview

Computational sequence analysis revisited: new databases, software tools, and the research opportunities they engender

Mark S. Boguski

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bldg. 38A, Rm. 8S-810, 8600 Rockville Pike, Bethesda, MD 20894

Abstract The increasing quantity and complexity of sequences and structural data for proteins and nucleic acids create both problems and opportunities for biomedical researchers. Fortunately, a new generation of practical computer tools for data analysis and integrated information retrieval is emerging. Recent developments in fast database searching, multiple sequence alignment, and molecular modeling are discussed and windows-based, mouse-driven software for CD-ROM and network information retrieval are described. Each method is illustrated with a practical example pertinent to lipid research. In particular, the connection among cholesteryl ester transfer protein, bactericidal permeability-increasing protein, and lipopolysaccharide-binding proteins is determined; novel repetitive sequence motifs in mammalian farnesyltransferase subunits and related yeast prenyltransferases are derived; biochemical insights from a three-dimensional model of human apolipoprotein D based on two insect lipocalins are discussed; the relationship between apolipoprotein D and gross cystic disease fluid protein from human breast is reviewed; and prospects for modeling apolipoprotein E-related proteins are described. In addition, information on a number of general and special-purpose sequence, motif, and structural databases is included.—**Boguski, M. S.** Computational sequence analysis revisited: new databases, software tools, and the research opportunities they engender. *J. Lipid Res.* 1992. **33:** 957-974.

Supplementary key words information retrieval • molecular modeling • motifs • cholesteryl ester transfer protein • prenylation • farnesyltransferase • apolipophorin III • apolipoprotein E • apolipoprotein D • lipocalin superfamily

CONTENTS

INTRODUCTION

Six years ago we reviewed, for the *Journal of Lipid Research*, the state of computational biology as applied to the analysis of molecular sequence data (1). Since that time, the number of available sequences has been doubling about every 22 months and software tools for information retrieval and analysis are achieving a new level of sophistication, integration, and accessibility. In parallel, detailed crystal structures of a number of proteins important in lipid biochemistry and metabolism have been elucidated. This convergence of new computer methods, greatly expanded sequence data, and new structural data have created unprecedented opportunities for molecular modeling and mutagenesis design that will advance structural analysis a quantum leap beyond the rather limited assessment of structure potential that has successfully dominated sequence analysis in this field for almost two decades.

The present review is not meant to be comprehensive but rather will concisely describe recent developments in the following areas: integration of sequence and bibliographic databases, rapid sequence similarity search tech-

nology, and multiple sequence alignment for detecting sequence "motifs" and as a prelude to tertiary structure modeling and design of mutagenesis experiments. As each method is introduced, I show how it can be used to analyze specific sequences including cholesteryl ester transfer protein (CETP), prenyl-protein transferases, apoE-related proteins, and the lipocalin superfamily.

INFORMATION RESOURCES

Molecular sequence data is the common currency of modern biomedical research and often provides exciting and unexpected links between diverse systems that accelerate research progress. Today one cannot afford to be unfamiliar with the sources and uses of DNA and protein sequence collections and much effort is currently being expended to make these data more accessible to the general scientific community (see below). A detailed discussion of sequence databases is beyond the scope of this review and more information may be found in several recent books (2–4). However characteristics of some major databases that are important for molecular and structural biology are presented in **Table 1**.

An exciting new source of sequence data has recently arisen in the context of genome research. In an attempt to produce a comprehensive survey of all expressed genes in particular tissues or organisms, a number of groups have embarked on projects that use robotics and automated sequencing technology to rapidly obtain partial sequence data on randomly selected clones from cDNA libraries (5). These partial cDNAs, also known as Expressed Sequence Tags or ESTs, have many potential uses, not the least of which is to accelerate the cloning of human genes for which homologs in other organisms are already known. One project has already identified more than two thousand genes from human hippocampus (6) with up to 20,000 additional sequences anticipated by the end of 1992 (J. C. Venter, personal communication). Similar work in *Caenorhabditis elegans* (a nematode worm that is a model organism for genome research) has already identified and mapped 1200 of the approximately 15,000 genes in this organism (7). A number of proteins pertinent to lipid research have already been identified in both the human and *C. elegans* EST collections (see below).

One of the problems with traditional sequence databases is that they represent large independent, yet considerably redundant, sources of information. Navigating through this "information space" can be so difficult and time-consuming that most laboratory-based scientists rarely, if ever, attempt it. Even when they do, they may be searching out-of-date collections with software that may be inappropriate for the particular questions they are ask-

TABLE 1. Some major databases for molecular and structural biology

Name	Contents	Availability
Entrez:Sequences	Nucleic acid and protein sequences with MEDLINE abstracts from published articles and direct submission from authors	CD-ROM, Network info@ncbi.nlm.nih.gov
GenBank	Nucleic acid sequences from published articles and by direct submission from authors. Translations of coding regions separately available (GenPept)	Network, CD-ROM, magnetic tape gos@genbank.bio.net
EMBL	European Molecular Biology Laboratories—nucleic acid sequences from published articles and by direct submission from authors	Network, CD-ROM, magnetic tape datalib@embl-heidelberg.de
NBRF/PIR	National Biomedical Research Foundation/Protein Identification Resource—protein sequences with annotations from published articles	Network, CD-ROM, magnetic tape pirmail@gunbrf.bitnet
SWISS-PROT	Protein sequences with annotations from published articles and translations of coding regions from EMBL; companion database of protein motifs (PROSITE) cross-referenced to SWISS-PROT entries	Network, CD-ROM, magnetic tape bairoch@cmu.unige.ch datalib@embl-heidelberg.de
PDB	Protein Data Bank—protein and nucleic acid three-dimensional structures by direct submission from authors of crystallographic and NMR data as well as molecular models	Network, magnetic tape pdb@bnlchm.bitnet
CCSD	Complex Carbohydrate Structural Database—complex carbohydrate, glycolipid, and glycopeptide sequences and annotations from published articles	IBM PC floppy diskettes CarbBank@uga.bitnet

Current releases at the time of this writing include: GenBank 70.0 (77,337,678 bases in 58,952 loci), NBRF/PIR 30.0 (9,697,617 residues in 58,952 sequences), SWISS-PROT 20.0 (7,500,086 residues in 20,654 sequences), PDB (approx. 800 coordinate sets with an additional 400 awaiting release), and CCSD (approx. 5,000 entries). To provide some idea of the redundancy among PIR, SWISS-PROT, and translated GenBank (GenPept), a composite NCBI collection of merged databases with exact sequence matches removed currently contains over 16 million residues in approximately 58,000 sequences. A number of other databases and analysis programs are freely available by anonymous file transfer protocol (ftp) from ncbi.nlm.nih.gov.

[a]As of October 1992, NCBI will assume responsibility for GenBank distribution.

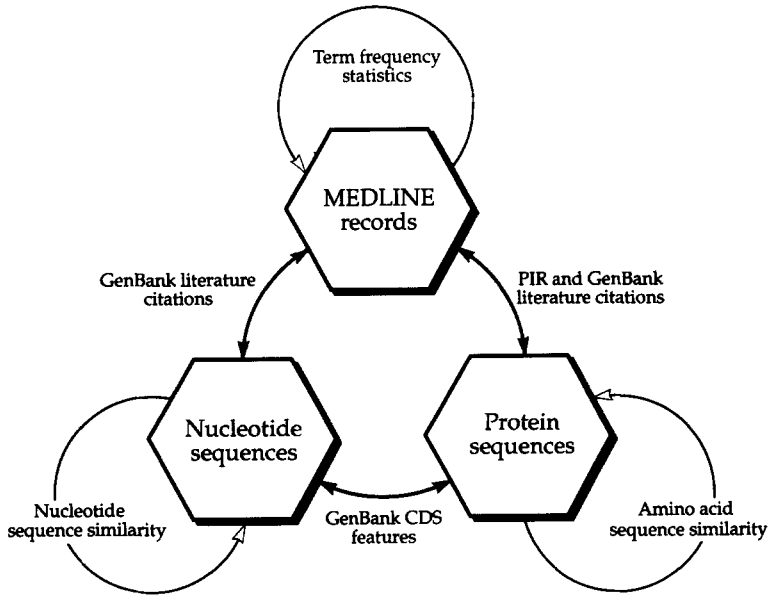

Fig. 1. Schematic diagram of *Entrez* databases. The beta test release CD-ROM contains over 86,000 MEDLINE citations with abstracts that have been indexed under the MeSH heading "molecular sequence data" or that have been cited in the sequence databases. More than 53,000 protein sequences and 35,000 nucleotide sequences from PIR and GenBank (see Table 1) are also present (additional sources of data will be included in future releases). These various data sets are linked by both explicit and implicit (computed) attributes as described in the text. These linkages are designed for integrated information retrieval as illustrated in Figs. 2–5.

ing. And even when something of potential interest is found, it often requires a trip to the library to seek out additional information that is critical to assessing biological significance but that is only present in the publication describing the sequence.

For the past several years, the National Center for Biotechnology Information (NCBI) has been developing information resources that integrate molecular sequence and structural data with the MEDLINE® bibliographic database (8). These resources are designed for easy access over national high-speed computer networks (see below) or via CD-ROM[1] on a personal computer. These systems are currently undergoing final testing and will be generally available by the time this article appears. I illustrate the use of NCBI's CD-ROM in the following section. Network-based applications are described later.

[1]CD-ROM is an acronym for "compact disc read-only memory" and is an optical memory system that can store any type of digitized data. CD-ROMs have been in use for distributing sequence databases for several years and a number of scientific journals have begun distributing back issues in this format. The advantage of CD-ROM is the enormous storage capacity. A typical hard disk in a desktop computer holds 80 megabytes (or 80 million characters) of data. Current technology permits the storage of about 660 megabytes on a CD-ROM. For comparison, the size of the entire human genome is estimated at 3,000 megabytes (3×10^9 nucleotides).

INTEGRATED INFORMATION RETRIEVAL

Entrez:Sequences is an information retrieval system that provides a common interface to nucleotide and protein sequence databases, the biomedical literature, and many of the explicit and implicit linkages that connect them (**Fig. 1**). On an Apple Macintosh® or IBM®-compatible computer (running Microsoft Windows®), or on a Unix® workstation, one can rapidly search several hundred million characters of sequence data and MEDLINE entries with just a few clicks of a computer "mouse."[2] Sequences and the publications that describe them not only have explicit links to each other but also implicit or emergent links to related sequences or articles that are determined by precomputed sequence "homologies" and term frequency statistics derived from titles, abstracts, and Medical Subject (MeSH®) Headings (Fig. 1). These linked entries are hereafter referred to as "neighbors." I illustrate the use of this system by an example.

Cholesteryl Ester Transfer Protein

Suppose one was interested in finding out something about cholesteryl ester transfer protein (9). One starts

[2]The term "mouse" is used to refer to the mechanical pointing devices in common use on many personal computers and workstations.

Entrez, types these terms into a text box, and presses the "Accept" button to retrieve articles that contain these terms (**Fig. 2**). After inspecting the list of first authors, years, and titles of the retrieved articles to determine which are most relevant to our present interests, one can choose *1)* to read the abstracts by "double clicking" on the titles, *2)* to search for related articles ("neighbors") among MEDLINE records by clicking the left hand check box, or *3)* to find corresponding sequence records by clicking the check box and then selecting a sequence database. When the latter procedure is carried out, one finds that there are three entries for CETP in the protein database (**Fig. 3**). The full text of these database records can then be viewed (Fig. 3).

For every new release of *Entrez*, the sequence databases have been compared against themselves (using the "BLAST" program, see below) and all significant similarities between sequence pairs are stored on the CD-ROM. These precomputed database search results constitute emergent links, some already known and some as yet undiscovered, among the sequences. One can, in effect, carry out extremely rapid homology searches against all known sequences by taking advantage of the stored results. Returning to our example, we see that CETP has twelve "neighbors" or homologs in the protein sequence database (**Fig. 4**). Some of these are just independent determinations of the sequence published by different groups, but we also see that CETP is related to a bactericidal permeability-increasing protein and lipopolysaccharide-binding protein. To further investigate these relationships, we can use the sequence records to link back to MEDLINE and retrieve the abstracts for inspection (**Fig. 5**).

In this case, the relationship among CETP and the other two proteins was already known (10, 11). However, one can easily imagine discovering and investigating new relationships using these tools.

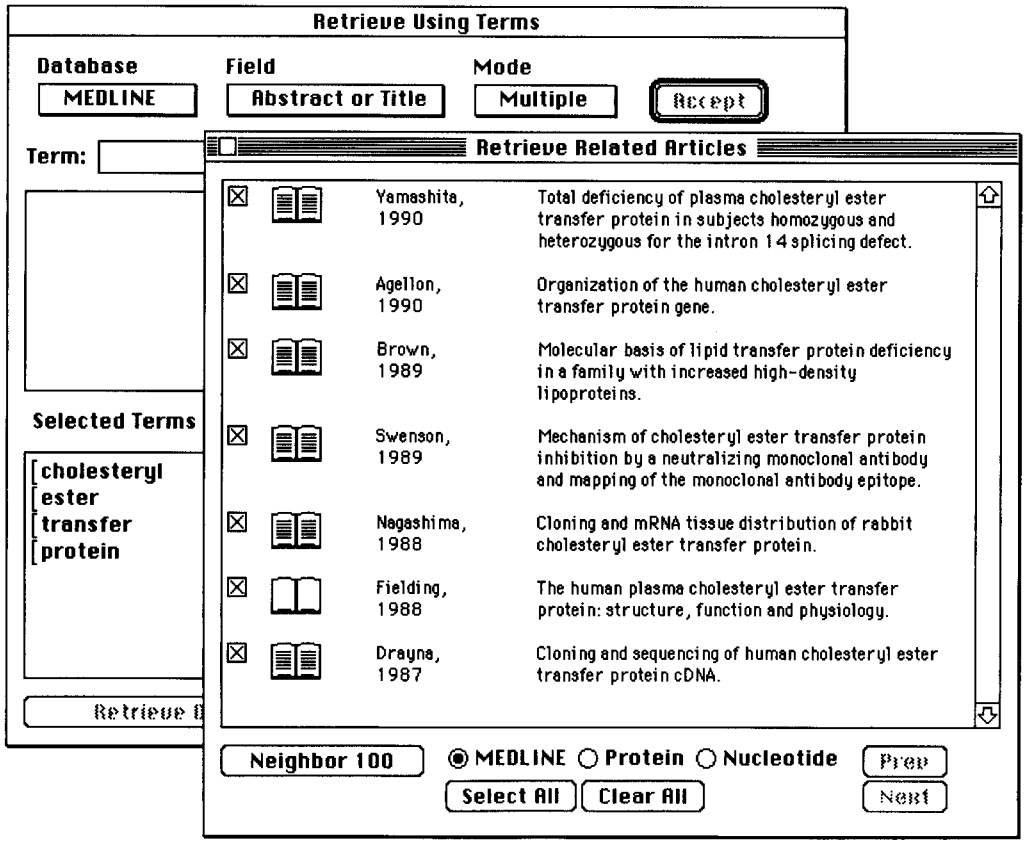

Fig. 2. Beginning an *Entrez* search. One typically selects an initial set of database records through a keyword query. In this case the database is MEDLINE and the keyword query string is "cholesteryl ester transfer protein." Seven MEDLINE records selected by this query are shown in the foreground screen image (Retrieve Related Articles) as a list of first authors, publication years, and article titles. Note that these seven articles are cross-referenced to 100 related articles (neighbors) in the MEDLINE database. One can view the full text of a MEDLINE abstract by "double-clicking" on the article title (see Fig. 5). Protein sequences to which these seven articles refer are retrieved by selecting the "Protein" button at the bottom of the screen. The searches depicted in Figs. 2–5 were performed on an Apple Macintosh® computer and the figures represent unedited "screen dumps" obtained during the search session. *Entrez* is also available for Microsoft Windows® and X-Windows/Unix® systems. Information on Entrez subscriptions (six updates per year for $57) and a detailed tutorial users manual are available on request from NCBI.

SEQUENCE SIMILARITY SEARCHING

It is often necessary, of course, to search the existing sequence collections for homologies to new, previously unpublished sequences and thus *Entrez* also includes a personal computer implementation of the BLAST program (12) for this purpose. But BLAST (for Basic Local Alignment Search Tool) is also implemented as a network service that permits easy access to very fast, multiprocessor computers and daily updates of the sequence collections.

Internet is a U.S. computer network that connects most major colleges, universities, research institutions, and government laboratories (13, 14). Many users of personal computers have access to Internet via local area networks. The BLAST network service allows one to formulate a database search locally, have it transparently migrate perhaps thousands of miles over Internet to a remote computer, and then return the results in a matter of seconds to the local system for inspection.

The BLAST program achieves at least an order of magnitude increase in speed over earlier database search programs and has the added advantage that a new theory for estimating the statistical significance of sequence similarities is applied to the results obtained (12). For reasons partly having to do with our new understanding of protein sequences as modular constructs, BLAST does not attempt to make global (end-to-end) alignments of sequences but rather focuses on identifying all significant local similarities among a pair of sequences. The use of this new tool and the advantages of this approach are illustrated in the following example.

Protein Prenyltransferases

The posttranslational attachment of isoprenoid lipids (or prenyl groups) to a number of regulatory proteins and oncogene products is necessary for their targeting and anchoring to various cellular membranes that are the sites of their action (15–18). Prenyl groups, such as farnesyl and geranylgeranyl pyrophosphates, are derived from the cholesterol biosynthetic pathway and become linked to specific target sites at the COOH-termini of various proteins via thioether bonds. These reactions are catalyzed by hetero-

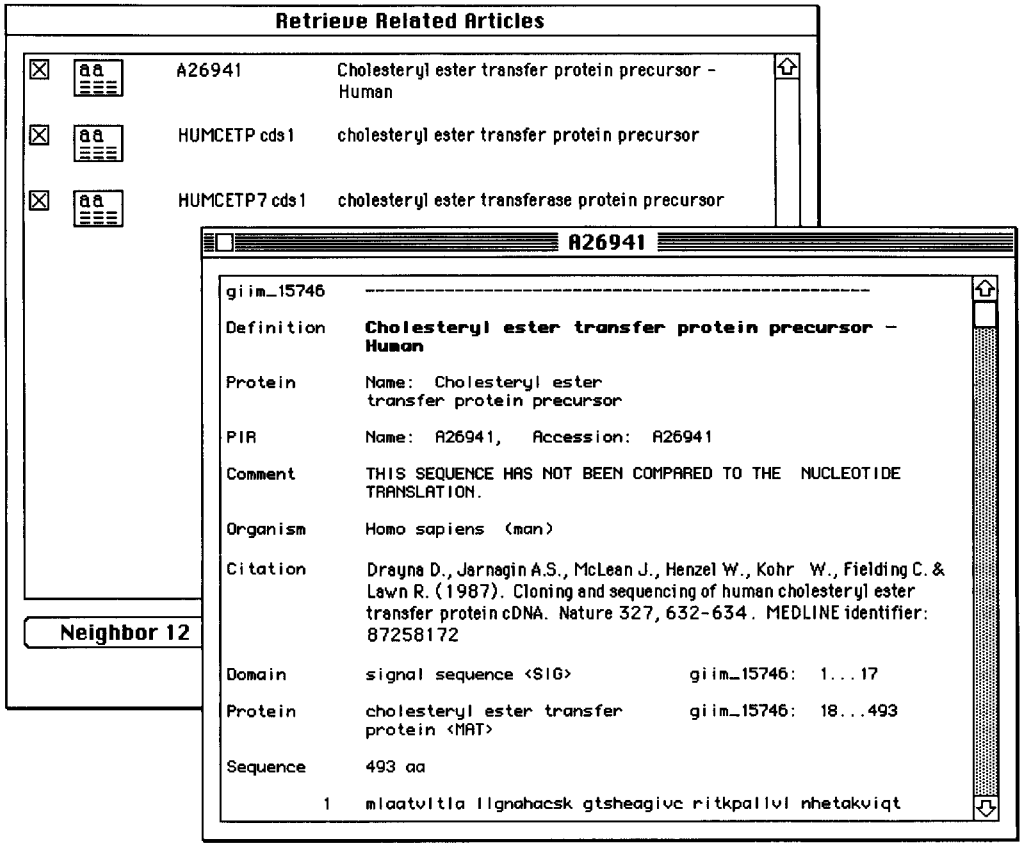

Fig. 3. Retrieving cross-referenced protein sequences. Sequence database entries corresponding to the seven published articles in Fig. 2 are shown in the background screen image as a list of accession numbers and sequence-organism names. By "double-clicking" on the first sequence name in this list, one pulls up the full text of one database record for human CETP as shown in the foreground screen image.

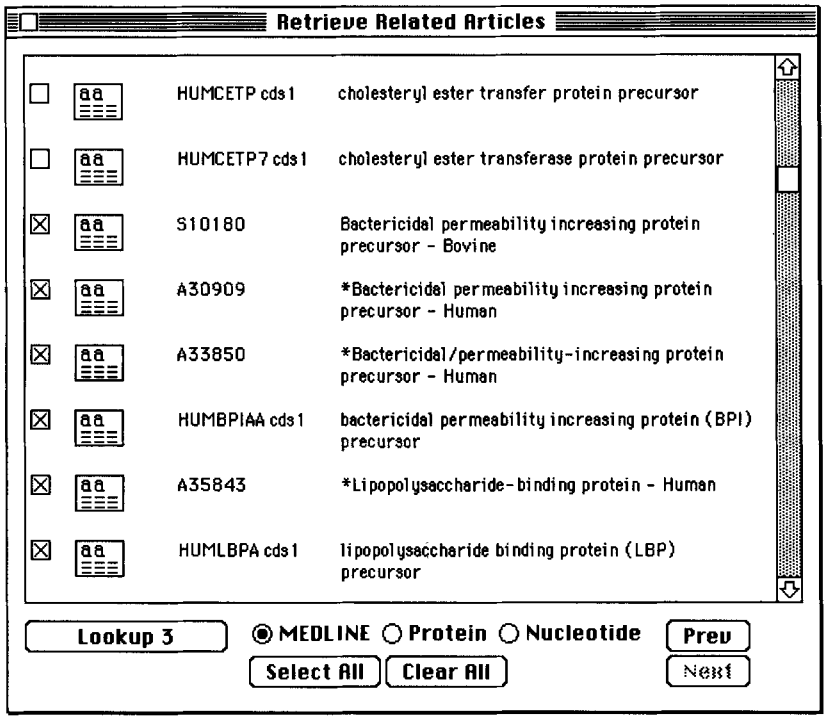

Fig. 4. Identifying homologous sequences. By pressing the button labeled "Neighbor 12" on the previous screen (Fig. 3), one accesses precomputed similarity search results to identify those sequences that are homologous to CETP. Due to the nature of these particular sequence databases, the list of related sequences includes some self-comparisons and also matches to multiple, semi-redundant entries for other proteins. One selects a subset of entries for further analysis using the computer "mouse" to place X's in the check boxes at the left-most portion of the window.

dimeric enzymes consisting of α and β subunits (19). The best-studied of these enzymes is mammalian farnesyltransferase (19, 20) although similar proteins are also present in yeast (21). Recently, cDNA sequences for a number of α and β subunits have been cloned (21–27).

Using the BLAST program with rat farnesyltransferase α subunit (hereafter FT-α) as a query sequence, I searched a comprehensive, composite collection of over 57,000 sequences[3] containing more than 16 million residues and the results are shown in **Fig. 6.** Note the extremely high scores (and low *P*-values) for the self-comparison and the match to the bovine homolog of FT-α (Fig. 6A). The "N" value in the right-most column (Fig. 6A) signifies that one "local" similarity was found between the query and matched sequences: in this case the "local" alignment is actually the same as a global alignment because the rat and bovine sequences are highly-conserved and no insertion/deletion mutations have occurred (Fig. 6B). From the fourth match down (*Xenopus lamin*, Fig. 6A), the results are not statistically significant. However the third match of FT-α to the yeast MAD2 gene product has a chance probability of only 4×10^{-16} indicating that these sequences are "homologous" in the evolutionary sense, i.e., they are derived from a common ancestral gene and may possibly have similar functions. Note in this case that N = 5 (Fig. 6A), indicating the presence of five significant local similarities between FT-α and Mad2 (shown in Fig. 6C). When one observes sequence homologies broken up into multiple, ungapped segments like this, it can mean: *1)* that sequence conservation has been nonuniform (with or without the presence of insertion/deletion mutations); *2)* that the sequences are modular and/or mosaic in structure; and/or *3)* that they have undergone intrasequence transpositions with respect to one another as in some cases of proteins containing internal repeats. Furthermore, intense but very localized similarities may even point to the existence of an enzyme active site or other functional motif (28). One can investigate these possibilities further using multiple sequence alignment techniques, which are the subject of the next section (see also ref. (29)).

Table 2 summarizes the BLAST database search results using the rat FT-β sequence as the query. Note that the magnitudes of the scores among the various yeast

[3]Contrast this value with the size of the database (only 3,712 sequences) used in our 1986 review (1). The current situation represents a 15-fold increase in the amount of data to be searched, but hardware and software advances have, at the same time, reduced the search time from minutes to seconds.

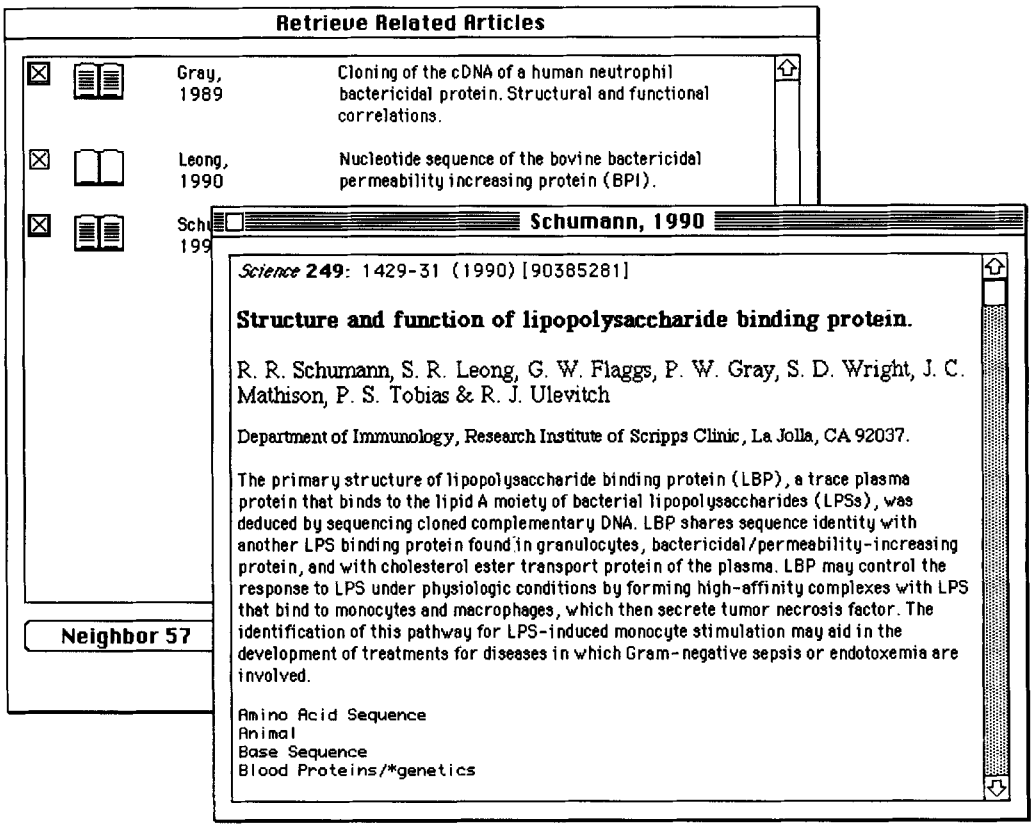

Fig. 5 Linking back to the scientific literature. The six selected sequence database records of Fig. 4 are cross-referenced back to MEDLINE by their bibliographic citations and it is seen that these sequences are described in three publications listed in the background screen image. The foreground screen image shows the full text of the MEDLINE record retrieved by "double-clicking" on the third title.

and mammalian homologs are similar to that of mammalian FT-α and yeast Mad2 (Fig. 6A). Also note the multiple, segmental nature of the local alignments (last column in Table 2), again similar to the situation with FT-α and Mad2.

In the previous section on Information Resources, I briefly described a new type of data library containing partial cDNA sequences or ESTs. For several reasons, it is necessary to adopt a somewhat different approach to sequence similarity searching here. The object of the search is to determine whether homologs of known proteins are present in an EST collection and, for reasons discussed in our 1986 review (1), it is always more sensitive to do this by comparing protein rather than DNA sequences. But because of the way in which the EST libraries are constructed, it is not known in advance whether the sequences are derived from the nontranslated or protein-coding regions of an mRNA, nor in the latter case is the correct translation reading frame known. Thus one must conceptually translate the EST sequences in all six reading frames (three each on the two complementary strands) and then compare each translation with the amino acid query sequence. A variant of the BLAST program, called TBLASTN, has been developed for this purpose (30) (W. Gish and D. J. States, unpublished results). Using this approach, one is able to identify homologs of the rat farnesyltransferase α subunit in both the human brain (6) and nematode collections (7).[4] Furthermore, the *C. elegans* library also contains a β subunit homolog most closely related to the yeast gene BET2 (Table 2).[4]

MULTIPLE SEQUENCE ALIGNMENT

Due to the exponential accumulation of new sequence data, it is now often the case that a sequence one is interested in has several known homologs. The alignment of a family of sequences is almost always more informative than alignments of any of the pairs (31). Using multiple sequences may improve the accuracy of the alignment and

[4]The database code for the human FT-α sequence is EST00211 (6) and, as for all clones in this library, the physical DNA can be ordered directly from the American Type Culture Collection (12301 Parklawn Dr., Rockville, MD 20852). The database codes for the *C. elegans* FT-α and Bet2 homologs are cm2e12 and cm20c6, respectively (7).

```
National Center for Biotechnology Information

Experimental GENINFO(R) BLAST Network Service

Databases available:

  nr          Non-redundant PIR+SwissProt+GenPept+GPUpdate, updated daily
              for efficient, complete searches of all four component databases
  pir         PIR 30.0 (complete), September 30, 1991
  swissprot   SWISS-PROT 20.0, December 1991
  genpept     GenPept (translated GenBank) Release 70.0, December 15, 1991
  gpupdate    daily GenPept update
  tfdaa       transcription factor (amino acid) database Release 4, Dec. 1991

  genbank     GenBank(R) Release 70.0 (complete), December 15, 1991
  gbupdate    daily GenBank(R) update

BLASTP 1.2.1MP [26-Dec-91]

Reference:  Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers,
and David J. Lipman (1990).  Basic local alignment search tool.  J. Mol. Biol.
215:403-410.

Query=  rat FT-alpha (residues 51-377)

Database:  Non-redundant PIR+SwissProt+GenPept+GPUpdate+BackBone, 3:35 AM EST
           Jan 2, 1992
           57,120 sequences; 16,082,577 total residues.

                                                                    Smallest
                                                                    Poisson
                                                          High  Probability
Sequences producing High-scoring Segment Pairs:          Score  P(N)       N

GPU:RATPFAS_1 CDS 57..1190 /product="farnesyltransferase ...  1803  1.4e-255  1
GP:BOVFTASE_1 Cow farnesyl-protein transferase alpha-subu...  1731  3.7e-245  1
GP:YSCMAD2_1  S.cerevisiae mitotic MAD2 gene, complete cd...    93  4.0e-16   5
SP:LAM2$XENLA LAMIN L(II) >GP:XELLAMLII_1 Xenopus laevis ...    66  0.22      1
PIR:S03966    Dystrophin-related protein - Human (fragmen...    59  0.92      1
SP:LAM2$MOUSE LAMIN B2 >GP:MUSLAMIN2_1 Mouse mRNA for lam...    55  0.93      2
GPU:EHDDVL3_1 CDS      18..2717 /standard_name="VP3" /pro...    57  0.99      1
PIR:ZEBP4L    Ea47 protein - Bacteriophage lambda | 2475....    56  1.0       1
PIR:XYRTFA    Fatty-acid synthase - Rat #EC-number 2.3.1....    55  1.0       1
SP:FAS$RAT    FATTY ACID SYNTHASE (EC 2.3.1.85) (CONTAINS...    55  1.0       1


Time to search database:  13.13u 0.33s 13.46t
Total cpu time:  13.44u 0.54s 13.98t
```

A

```
>GPU:RATPFAS_1 CDS      57..1190 /product="farnesyltransferase alpha subunit"
            /gene="farnesyltransferase alpha subunit" /codon_start=1
            Length = 377

 Score = 1803 (867.9 bits), Expect = 1.4e-255, P = 1.4e-255
 Identities = 327/327 (100%), Positives = 327/327 (100%)

                    (self alignment omitted)

>GP:BOVFTASE_1 Cow farnesyl-protein transferase alpha-subunit mRNA, complete
            cds.
            Length = 329

 Score = 1731 (833.3 bits), Expect = 3.7e-245, P = 3.7e-245
 Identities = 310/327 (94%), Positives = 322/327 (98%)

Query:    51 MDDGFLSLDSPTYVLYRDRAEWADIDPVPQNDGPSPVVQIIYSEKFRDVYDYFRAVLQRD 110
             MDDGFLSLDSPTYVLYRDR+EWADIDPVPQNDGP+PVVQIIYSEKF+DVYDYFRAVLQRD
Sbjct:     1 MDDGFLSLDSPTYVLYRDRPEWADIDPVPQNDGPNPVVQIIYSEKFQDVYDYFRAVLQRD 60

Query:   111 ERSERAFKLTRDAIELNAANYTVWHFRRVLLRSLQKDLQEEMNYIIAIIEEQPKNYQVWH 170
             ERSERAFKLTRDAIELNAANYTVWHFRRVLL+SLQKDL+EEMNYI AIIEEQPKNYQVWH
Sbjct:    61 ERSERAFKLTRDAIELNAANYTVWHFRRVLLKSLQKDLHEEMNYISAIIEEQPKNYQVWH 120

Query:   171 HRRVLVEWLKDPSQELEFIADILNQDAKNYHAWQHRQWVIQEFRLWDNELQYVDQLLKED 230
             HRRVLVEWL+DPSQELEFIADIL QDAKNYHAWQHRQWVIQEF+LWDNELQYVDQLLKED
Sbjct:   121 HRRVLVEWLRDPSQELEFIADILTQDAKNYHAWQHRQWVIQEFKLWDNELQYVDQLLKED 180

Query:   231 VRNNSVWNQRHFVISNTTGYSDRAVLEREVQYTLEMIKLVPHNESAWNYLKGILQDRGLS 290
             VRNNSVWNQR FVISNTTGY+DRA+LEREVQYTLEMIKLVPHNESAWNYLKGILQDRGLS
Sbjct:   181 VRNNSVWNQRYFVISNTTGYNDRAILEREVQYTLEMIKLVPHNESAWNYLKGILQDRGLS 240

Query:   291 RYPNLLNQLLDLQPSHSSPYLIAFLVDIYEDMLENQCDNKEDILNKALELCEILAKEKDT 350
             +YPNLLNQLLDLQPSHSSPYLIAFLVDIYEDMLENQCDNKEDILNKALELCEILAKEKDT
Sbjct:   241 KYPNLLNQLLDLQPSHSSPYLIAFLVDIYEDMLENQCDNKEDILNKALELCEILAKEKDT 300

Query:   351 IRKEYWRYIGRSLQSKHSRESDIPASV 377
             IRKEYWRYIGRSLQSKHS ESD P++V
Sbjct:   301 IRKEYWRYIGRSLQSKHSTESDPPTNV 327
```

B

```
>GP:YSCMAD2_1 S.cerevisiae mitotic MAD2 gene, complete cds. >BB:B48623 [Peptide
            290 aa] putative calcium-binding protein [Saccharomyces cerevisiae]
            Length = 290

 Score = 93 (44.8 bits), Expect = 2.9e-05, P = 2.9e-05
 Identities = 20/65 (30%), Positives = 38/65 (58%)

Query:   214 RLWDNELQYVDQLLKEDVRNNSVWNQRHFVISNTTGYSDRAVLEREVQYTLEMIKLVPHN 273
             ++W+ EL  V+ LL +D RN   W+ R+ V++N  + +++++   E +Y    I+    N
Sbjct:    85 KVWQTELAVVNKLLEQDARNYHGWHYRRIVVGNIESITNKSLDKEEFEYPTIKINNNISN 144

Query:   274 ESAWN 278
              SAW+
Sbjct:   145 YSAWH 149

 Score = 70 (33.7 bits), Expect = 0.064, Poisson P(2) = 3.7e-08
 Identities = 13/35 (37%), Positives = 21/35 (60%)

Query:   212 EFRLWDNELQYVDQLLKEDVRNNSVWNQRHFVISN 246
             E   WD+EL +V  LLK+  +   +WN+R  V+ +
Sbjct:    44 EIPFWDKELVFVMMLLKDYPKVYWIWNHRLWVLKH 78

 Score = 66 (31.8 bits), Expect = 0.24, Poisson P(3) = 7.9e-13
 Identities = 11/29 (37%), Positives = 18/29 (62%)

Query:   185 ELEFIADILNQDAKNYHAWQHRQWVIQEF 213
             EL F+  +L++ +K Y  W HR WV+  +
Sbjct:    51 ELVFVMMLLKDYPKVYWIWNHRLWVLKHY 79

 Score = 54 (26.0 bits), Expect = 14., Poisson P(4) = 8.3e-12
 Identities = 9/30 (30%), Positives = 17/30 (56%)

Query:   150 EEMNYIIAIIEEQPKNYQVWHHRRVLVEWL 179
             EE +Y    I++   NY  WH+R  ++ ++
Sbjct:   129 EEFEYPTIKINNNISNYSAWHQRVQIISRM 158

 Score = 54 (26.0 bits), Expect = 14., Poisson P(5) = 4.0e-16
 Identities = 11/34 (32%), Positives = 22/34 (64%)

Query:   116 AFKLTRDAIELNAANYTVWHFRRVLLRSLQKDLQ 149
             A K T + +E N+   ++W++RR ++ SL  +L+
Sbjct:    11 ALKKTSELLEKNPEFNAIWNYRRDIIASLASELE 44
```

C

Fig. 6. BLAST database search results using the Rat FT-α sequence as the query. Output of the BLAST program is shown in three parts. A. Header information and "hit list" of sequence matches. The upper section displays the databases available for searching, individually or as a comprehensive, composite collection from which exact duplicates have been excluded (see also Table 1). Rat FT-α is identified as the query sequence and the nonredundant database has been selected for searching. Although the full-length sequence of rat FT-α contains 377 residues (24), the first 50 residues includes an oligoproline segment as well as alanine and glycine di- and tripeptides that result in hundreds of spurious matches to other sequences in the database that possess similarly biased amino acid compositions. To avoid these false-positive matches, only residues 51–377 of rat FT-α were used for searching. The lower section contains one-line summaries of the search results that include *1)* a code for the parent database separated by a colon from the sequence retrieval key, *2)* a text string of descriptive information including the sequence name, and *3)* summary statistics consisting of the raw score, *P*-value and number of segments matched. So, for example, GPU:RATPFAS_1 means that a match was found in the "GenPept update" database and that the code for the matched sequence is RATPFAS_1. Note the entire search of more than 16 million residues required only 13.5 seconds. B. Significant local alignments. Each matching sequence is shown in alignment with the query sequence accompanied by a more expansive descriptive text string and more detailed statistics. The alignment of rat FT-α with itself is omitted. The alignment of rat and bovine FT-α is shown next. This alignment has a raw score of 1731, a 3.7×10^{-245} probability of chance occurrence and the two sequences are 94% identical. Conservative amino acid substitutions are indicated by "+" in the alignment. Because rat and bovine FT-α are so similar, the "local" alignment is the same as the "global" alignment. C. More significant local alignments. In this case, due in part to evolutionary divergence between the yeast and mammalian sequences, the similarity between rat FT-α and the yeast MAD2 gene product consists of a series of significant local alignments rather than an uninterrupted end-to-end alignment. Another contributing factor is that both sequences are composed of multiple, internal repeats as evidenced by the presence of numerous "intersecting" alignments among the various sequence pairs. (This phenomenon is discussed more fully later.) Although the raw scores for the matches are much lower than for the two mammalian homologs, the associated *P*-values in this case leave little doubt that yeast Mad2 is not related to FT-α by chance. The significance of this relationship is further supported by multiple alignment studies described in the subsequent section.

allows one to assess the range of sequence variation that is still compatible with function. Multiple alignment is important for the recognition of sequence patterns or motifs and can be of great assistance in structure prediction (32) and molecular modeling as described below (see also ref. 2).

Global versus Local Methods and Sequence Motifs

There are two major approaches to multiple alignment that can be used separately or in concert (e.g., refs. 29, 33). Global alignments attempt to match a group of sequences along their entire lengths and are most applicable to sequences that are relatively short, approximately equal in length, and have a colinear relationship (34). On the other hand, local alignment methods are best suited for longer sequences that vary considerably in length and may share only isolated regions of strong similarity separated by variable-length segments with little or no sequence conservation (e.g., refs. 33, 35). Local alignment is also the method of choice for analyzing sequences containing internal repeats (29, 31). Local alignment methods are computationally more efficient and the results are more amenable to statistical analysis (28).

"Motifs" are conserved patterns of amino acids or

TABLE 2. Results of database searching with the rat FT-β amino acid sequence

Related Sequence	Karlin Score	*P* Value	No. of Segments Matched
Rat FT-β (self-comparison)	2410	0	1
Yeast Ram1 (Dpr1)	193	5.0×10^{-31}	4
Yeast Bet2	117	4.9×10^{-11}	3
Yeast Cdc43 (Cal1)	82	3.0×10^{-4}	2

Conditions of the search were the same as in Fig. 6 except that the complete sequence of rat FT-β was used as the query. Retrieval codes for the sequences referred to in this table are as follows: rat FT-β (GenBank/GenPept locus RATFTBS__1), yeast RAM1/DPR1 gene product (NBRF/PIR code A30135), yeast BET2 gene product (GenInfo Backbone seqid 32756), and yeast CDC43/CAL1 gene product (GenBank/GenPept locus YSCCAL1__1).

nucleotides that are common to a group of functionally related sequences and often encode or specify the discrete function. For example, the pattern Gly-Asp-Ser-Gly-Gly represents the serine protease active site whereas purine nucleotide binding sites can be summarized by Gly-X-X-X-X-Gly-Lys-Ser/Thr where "X" may be any residue. For nucleotide sequences, there is the well-known "TATA" box that is required for accurate transcription initiation and the "Shine-Dalgarno" sequence that is necessary for optimal translation initiation in prokaryotes. Several hundred motifs have been compiled in printed form (36, 37), but machine-readable collections (38, 39) are essential for the purposes of database searching and automated motif detection.[5]

Motifs are almost always derived from a multiple alignment of a group of functionally related sequences. (More rarely they are defined by superimposing two or more three-dimensional structures.) Motifs are often presented as "consensus sequences" (1) but various quantitative or semi-quantitative representations have been devised (reviewed in refs. 4, 36). These are more useful for computerized searching and a number of commercial and public domain computer programs are available for this purpose (4, 36). Motifs vary tremendously in their specificity and predictive value (38) and the discovery of a particular pattern in a sequence should be interpreted with caution. Nevertheless, they are extremely useful for making functional inferences about a new sequence, even in the absence of global homology.

Application to Prenyltransferases

The Multiple Alignment Construction and Analysis Workbench (or MACAW) is a Microsoft Windows®-based system for easily and flexibly computing multiple alignments among a group of sequences (28). I used the MACAW program to study multiple alignments of FT-α and FT-β and related sequences and some of the results are shown in **Fig. 7**. MACAW allows one to simultaneously visualize sequence similarities as schematic block diagrams and also at the resolution of individual amino acid residues. One can also assess the statistical significance of various local alignments and interactively explore alternative alignments or guide the alignment process based upon external data, such as the experimentally determined location of disulfide bonds, etc.

The analysis shows that FT-α and it homologs contain a large, central domain that is composed of five internal repeats that are highly significant and conserved among the yeast and mammalian sequences (Fig. 7A). This repetitive, segmental nature of the sequence relationships was even evident in the database search results (see above) but is difficult to discern using more traditional global alignment techniques (25). An alignment of all 15 repeats from the three sequences is shown in **Fig. 8A** and they are seen to contain an invariant tryptophan and several other highly conserved residues.

Although they are somewhat more complex in structure and less conserved in sequence than FT-α repeats, repetitive motifs are also a predominant feature of FT-β and related sequences (Figs. 7B, 8B). Once again, strong indications of this phenomenon are present in the database search results (Table 2) and were overlooked when the sequences were globally aligned (26, 27). In the case of FT-β repeats, there is a conserved Cys-His-Cys motif (Fig. 8B) that may represent a novel type of zinc coordination domain (40): farnesyltransferase activity shows an absolute requirement for Zn^{2+}, probably at the peptide-binding site on the β subunit (41). Other structural and functional ramifications of these novel repetitive motifs are discussed elsewhere (29, 40).

Internal repeats are, of course, the predominant structural and functional feature of the apolipoprotein sequence family (1, 42, 43). However, as should be evident from the preceding example, intrasequence repeats are not a rare phenomenon and, indeed, internal repeats are estimated to be present in 10-20% of all protein sequences (31). Traditional dot-matrix algorithms used for this purpose have been superseded by more sophisticated

[5]Release 8.00 of the PROSITE database (Dec. 1991) contains 605 motifs or "signatures" and is available by electronic mail (info@ncbi.nlm.nih.gov). Release 4.0 the Transcription Factor Database currently contains 1887 "sites" and is also available by email (info@ncbi.nlm.nih.gov).

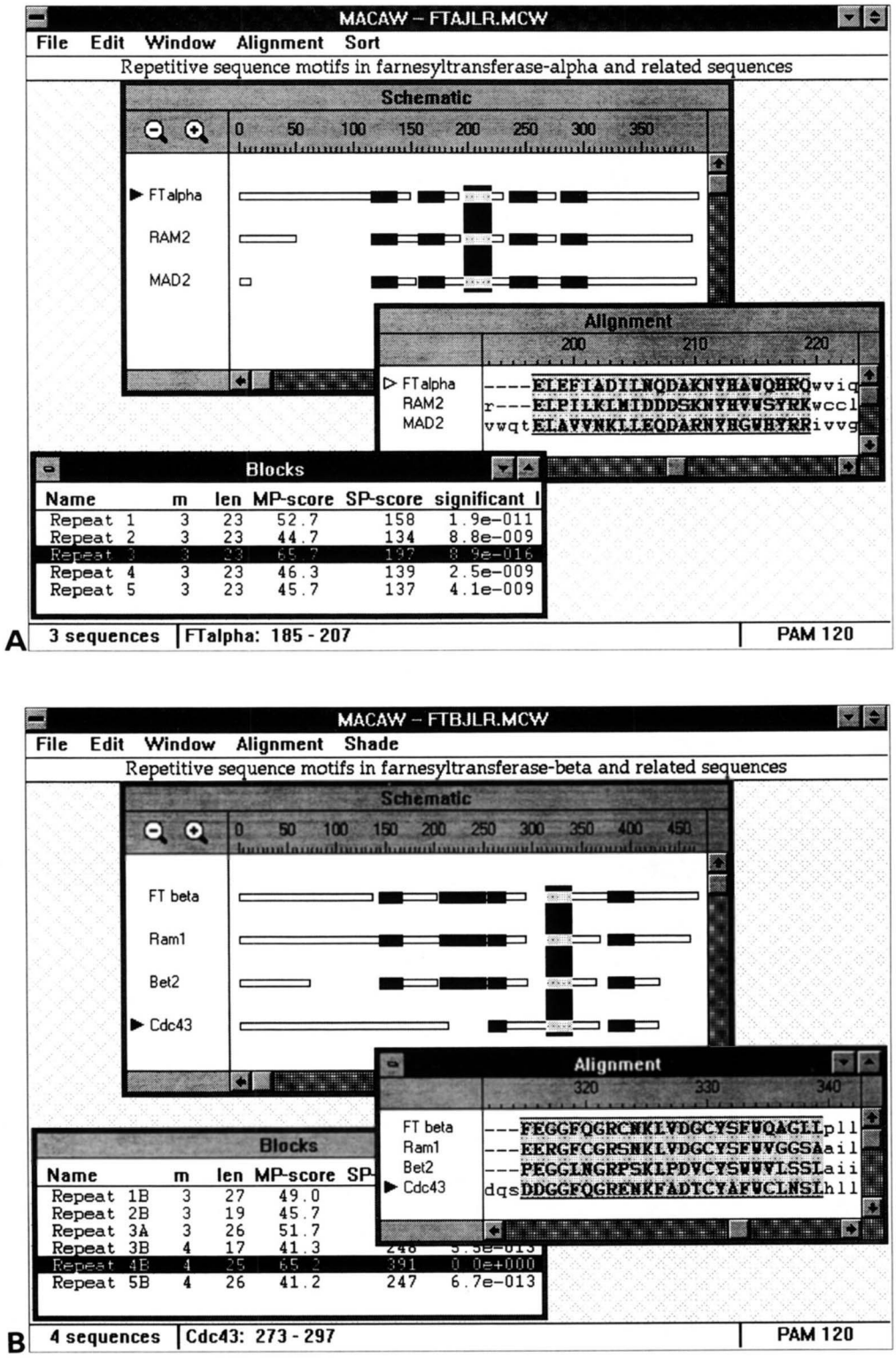

Fig. 7. Screen "snapshots" during multiple alignment analysis using the MACAW program. A. FT-α homologs. The locations and relative sizes of five internal repeats within each of the three sequences are shown in the "Schematic" window as a series of black rectangles on elongated white rectangles representing the full-length proteins. The third repeat is highlighted by a vertical black bar. A list of the repeats along with some associated scores and statistics are shown in the "Blocks" window. The "Alignment" window displays the local, three-way alignment representing repeat no. 3 and its precise location in the rat FT-α sequence (residues 185–207) is indicated in the lower left portion of the screen image. B. FT-β homologs. Note that, in contrast to FT-α repeats (panel A), FT-β repeats are more complex with regard to spacing and sequence conservation. In particular, the central domain of Cdc43 is not well conserved in comparison to the four-way alignment of C-terminal sequences. The α and β repeats are described more fully in Fig. 8. The analyses depicted here were performed on an IBM-compatible personal computer running Microsoft Windows.

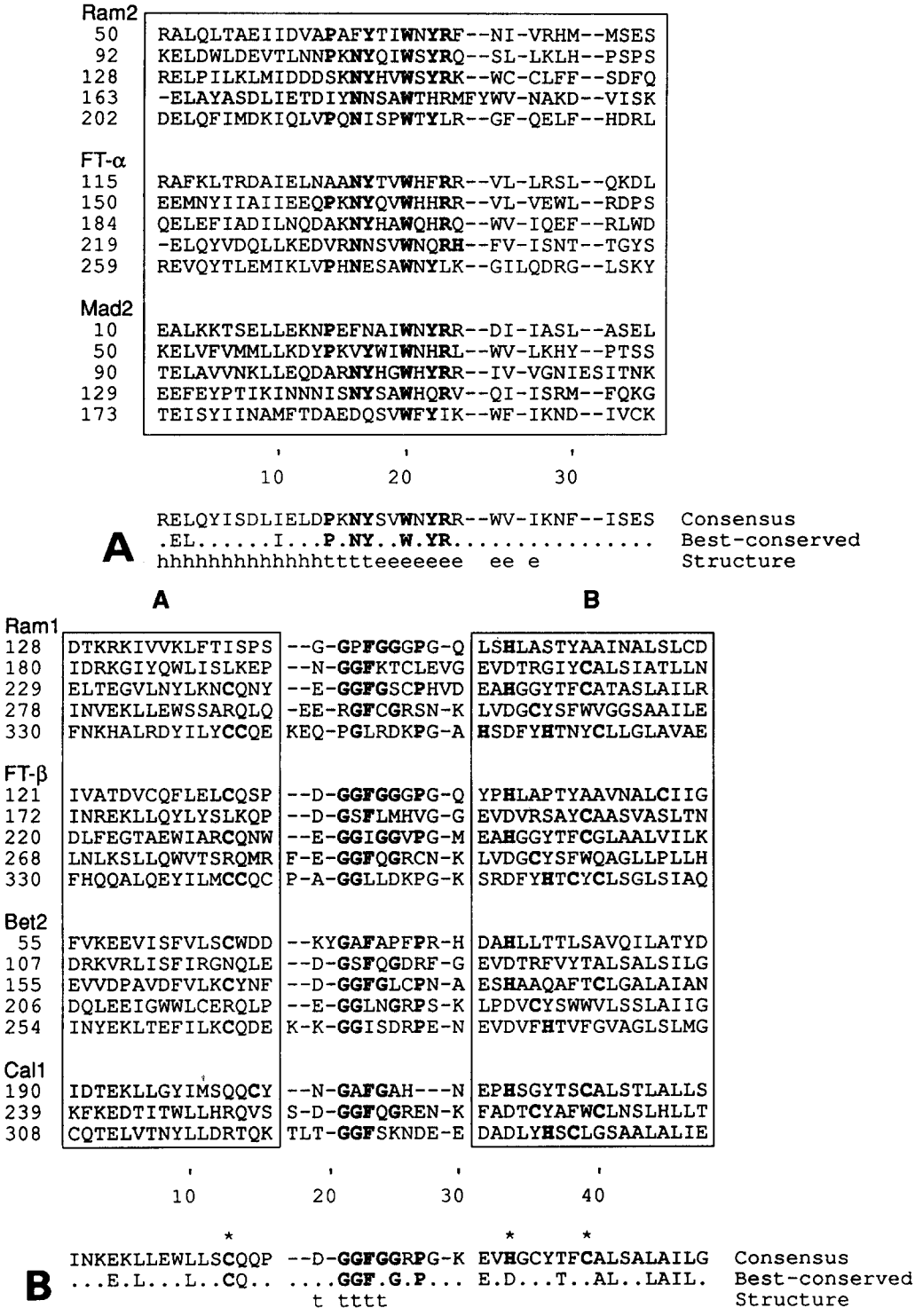

Fig. 8. Alignments of internal repeats from prenyltransferase subunits. Significant internal repeats were identified using the MACAW program (Fig. 7) and then extracted from the parent proteins along with variable flanking sequences and subjected to gapped multiple alignment (34). A. FT-α-related sequences. Fifteen repeats from three sequences are shown. (Bovine FT-α is excluded because it is 94% identical to rat FT-α (Fig. 6) and thus adds little to the analysis.) An ungapped, core block of 23–24 residues (see Fig. 7A) constitutes the most highly conserved unit but, allowing for small insertion/deletion mutations, the repeat is actually 34–35 residues in length. Residue numbers on the left indicate locations of the repeats within the parent proteins. Bold-faced type indicates conserved motifs as described in ref. 40. "Best-conserved" refers to residues that are identical in at least 45% of the repeats. Secondary structure prediction follows the evolutionary comparison strategy of Crawford et al. (32); "h" = α-helix, "t" = turn and "e" = β-strand. Functional implications of the repetitive tryptophan residues are discussed elsewhere (40). B. FT-β-related sequences. Eighteen repeats from four sequences are shown. In contrast to FT-α repeats (panel A), the average length of FT-β repeats is 44–45 residues and the central conserved motif is flanked by small gaps representing hypothetical insertion/deletion mutations. Asterisks mark the Cys-His-Cys motif that may be involved in zinc coordination.

methods (e.g., ref. 44) but graphic matrix plots remain a most useful method to visualize the results. "Dot plot" methods have recently been modernized (44) and adapted to view the results of multiple alignment computations (29). Prenyltransferase sequences have been analyzed using these techniques (29).

KNOWLEDGE-BASED OR COMPARATIVE HOMOLOGY MODELING

Ever since the amphipathic helix hypothesis was first annunciated nearly two decades ago (45), molecular modeling has been an important tool for studying the structure and function of proteins involved in lipid metabolism (43, 46). Modeling in this sense consists largely of predicting the secondary structure of peptides on a residue-by-residue basis (using probabilistic rules) and/or studying patterns of individual or averaged side-chain properties along the sequence (47, 48). This general approach has resulted in some fairly accurate correlations with subsequently determined structures. For example, four of the five amphipathic helices observed in the crystal structure of insect apolipophorin-III (49) were previously identified by Kanost et al. (50) using computational techniques for detecting amphipathic structures in proteins (51). Also, major characteristics of the structure of human apolipoprotein E (52) were anticipated by the analysis of structure potential of repetitive sequence motifs (1, 53, 54).

In contrast to this approach of predicting the secondary structure of proteins from a general knowledge of side chain properties and conformational propensities, it is now often possible to predict the tertiary structure of a protein by modeling it on the three-dimensional structure of an evolutionarily related sequence or homolog (reviewed in ref. 55). This technique, based on the availability of specific structures and homologous sequences, is referred to as "knowledge-based" (56) or "comparative homology" (57) modeling. Due to the steady accumulation of new structures, together with the rapid growth in sequence data, this method is being applied to an increasing number of protein families. For example, based on sequence similarity between retroviral proteases and the pepsin family of aspartyl proteases, an accurate model of the human immunodeficiency virus (HIV-1) protease was constructed (58) 2 years before its crystal structure was determined (59). Homology-based modeling capabilities are becoming increasingly common in commercial software packages for molecular mechanics and dynamics simulation.

For the lipid research field, several relevant crystal structures have become available in the past few years upon which one may predicate models of related proteins. Breiter et al. (49) have determined the molecular structure of an insect hemolymph lipid-transport protein, apoLp-III, which takes the form of an elongated helical bundle composed of five long amphipathic α helices. Interestingly, and despite the absence of significant similarity in primary sequence, the 22,000 M_r LDL receptor binding domain of human apoE is also an elongated helical bundle but composed of four amphipathic helices (52). The lack of evolutionary homology between apoLp-III and apoE indicates that any similarity in their three-dimensional structures is the result of convergent evolution.[6] In contrast, apoA-I and A-IV are recently diverged from apoE (42, 53) and it is obviously attractive to model aspects of apoA-I and A-IV structure based on the apoE coordinates (52) (see below).

Two other protein families (beside insect apoLp-III and human apoE) possess superficially similar topologies despite an absence of sequence similarity: the myelin P2 superfamily (that includes the intracellular fatty acid-binding proteins) and the lipocalin superfamily (that includes human serum retinol-binding protein and apolipoprotein D). The myelin P2/FABP structure is a "β-clam" composed of two five-stranded β sheets (60, 61); the lipocalin structure is also a β-clam but its two β sheets have only four strands each (**Fig. 9**) (62).

ApoD and the Lipocalin Superfamily

The lipocalin superfamily, whose general function is to bind small hydrophobic ligands for a variety of biological purposes, is represented by more than 40 proteins that fall into approximately 12 sequence families (63). Two of the most recently discovered members of this remarkably diverse family are prostaglandin D synthetase from human brain (the first lipocalin with identified enzymatic activity (63, 64), and crustacyanin, a protein whose function is to bind the carotenoid that imparts the blue coloration to the carapace of the European lobster (65).

For four lipocalins (insecticyanin, bilin-binding protein, retinol-binding protein, and β-lactoglobulin), high-resolution crystal structures have been determined (61, 66–69). Using database searching and multiple alignment analysis methods as described in previous sections (see also ref. 2), human apolipoprotein D was shown to be most closely related to insecticyanin and bilin-binding protein and a three-dimensional model of apoD was constructed (**Fig. 10**) based upon the atomic coordinates of insecticyanin (70). At least two novel insights were derived from detailed studies of this model of apoD: *1)* that cholesterol and/or cholesteryl esters were probably not the most likely ligands for apoD as had previously been conjectured (71); and *2)* that apoD's association with lipids and lipoproteins (72–74) may be a relatively nonspecific consequence of hydrophobic surface loops at the entrance

[6]Recent theoretical studies indicate that convergent evolution of similar structures may be more common than previously thought (83).

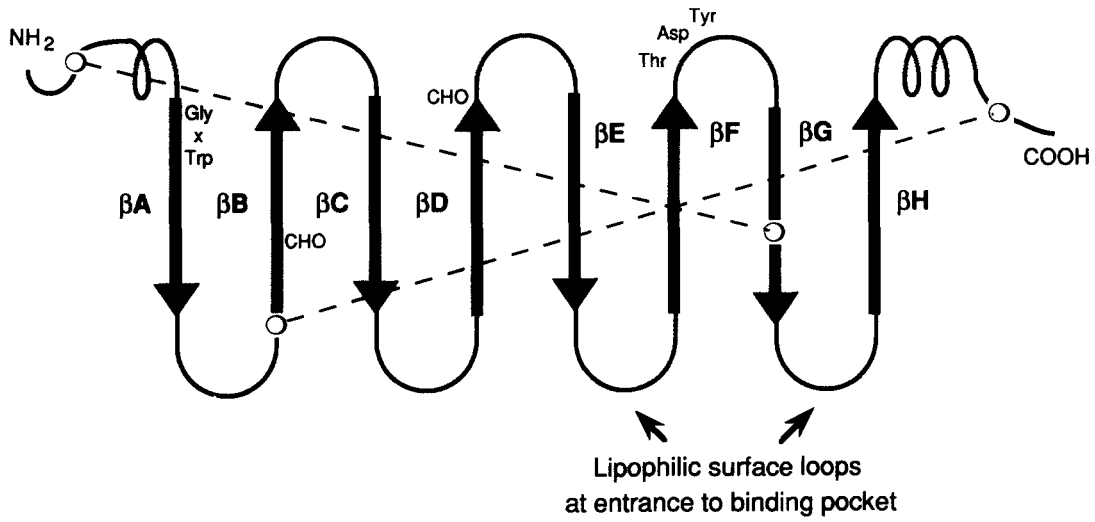

Fig. 9. Schematic diagram of human apoD. Eight antiparallel β strands are labeled βA through βH and are represented by arrows. These strands are connected by short loops or turns. Strands A–D and E–H form two orthogonal β sheets. A C-terminal α helix is indicated by a spiral. Two intersecting disulfide bonds (linking Cys^{8}–Cys^{114} and Cys^{41}–Cys^{165}) are shown as dashed lines. This intersecting pattern of disulfide bonds is characteristic of the apoD/insecticyanin subfamily of lipocalins (70). The tripeptides Gly^{24}–X–Trp^{26} and Thr^{103}–Asp^{104}–Tyr^{105} represent two sequence motifs that are almost universally conserved among all lipocalin superfamily members (2, 62). The approximate locations of two glycosylation sites are indicated by "CHO." Details of the lipophilic surface loops are given in Fig. 10.

to the binding pocket (70). This structural feature may also account for the low-affinity binding of certain lipophilic molecules (see below).

Although in preliminary experiments apoD was shown to bind bilirubin (but not cholesterol) in vitro (70), it may be the case that apoD has multiple physiologic ligands as for some other members of the lipocalin family, such as α_1-microglobulin which binds a wide variety of basic drugs and steroids (75). Of considerable interest in this regard is the recent demonstration that apoD is probably identical to the major protein component (GCDFP-24) in cyst fluid isolated from women with fibrocystic changes in their breasts (76). GCDFP-24 was originally characterized as a 24,000 M_r, progesterone-binding glycoprotein

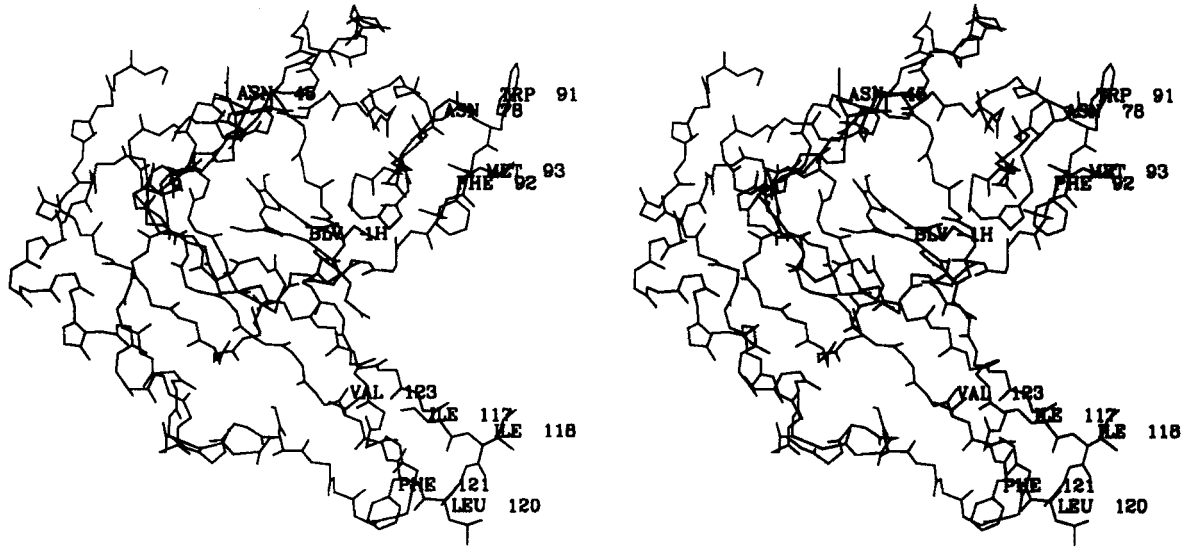

Fig. 10. Stereoscopic diagram of the structural model of human apoD with biliverdin ligand. The apoD model coordinates (identification code 1APD) were obtained directly from the Brookhaven Protein Data Bank (see Table 1). For the majority of the structure, only the protein backbone is shown but selected amino acid side chains are also displayed. In particular, solvent-accessible lipophilic residues that occur in surface loops at the entrance to the binding pocket are shown (Trp^{91}–Phe^{92}–Met^{93} and Ile^{117}–Ile^{118}, Leu^{120}–Phe^{121}, Val^{123}). The model suggests that this is the surface of apoD that interacts with HDL particles. Two potential N-linked glycosylation sites (Asn^{45}, Asn^{78}) are also shown. Note that carbohydrate moieties would point away from the lipophilic surface and thus would not be expected to sterically interfere with ligand binding or HDL association. A detailed description of binding pocket residues and other aspects of this structural model is found in ref. 70.

that exists as a tetramer in breast cyst fluid (77) and several steroid-binding studies have been carried out (78, 79). What these studies have demonstrated for GCDFP-24/apoD is low-affinity ($K \sim 10^{-6}$ l/mol) for progesterone which is maximal at nonphysiological pH (4–4.5) and with a stoichiometry indicative of only one/fourth molecule of ligand per monomer of protein (78). Competitive binding studies and determination of affinity constants under identical conditions would undoubtedly help to establish the relative importance of various apoD ligands.

Whatever these ligands may be, the nearly ubiquitous tissue distribution of apoD in mammals (71, 74, 80) indicates that it serves some essential function(s). Based on its ability to bind heme-related compounds, it has been suggested that apoD may play a role in free radical scavenging and/or porphyrin metabolism (70). Recent results may even point to unidentified enzymatic activities in apoD and other lipocalins (63).

Apolipoproteins E, A-I, and A-IV

Whatever the function(s) of apoD finally turn out to be, it is clear that its modeling led to new explanations for known properties and also to predictions about function that would have been difficult or time-consuming to infer by other means. It is enticing to imagine what new insights or explanations might be gained from modeling apolipoproteins A-I and A-IV based on the atomic coordinates of apoE. But are the sequences even similar enough to think modeling might be worthwhile? Results of Chothia and Lesk (81, 82) on the relationship between sequence and structure provide some encouragement. They showed that even down to a residue identity of ~20%, superimposed core structures of homologous proteins have a root mean square deviation of about 1.8 angstroms. Apolipoprotein D is 30% identical to insecticyanin upon which its model was based (70) and a useful model of bilin-binding protein was built on the coordinates of retinol-binding protein which is only 10% identical (66). In comparison, multiple alignment studies (M. Boguski, unpublished observations) show that the crystallized LDL receptor-binding domain of human apoE (residues 24–166) (52) is 26% identical to apoA-IV (residues 2–144) and 60% similar when conservative amino acid substitutions are taken into account. ApoE (residues 24–166) is about 18% identical (49% similar) to apoA-I (residues 12–154). This level of similarity seems sufficient to justify comparative homology modeling.

SUMMARY AND PERSPECTIVES

Sequences accrue, software advances, and the data-rich future of molecular biomedicine unfolds. In this review, I have tried to survey some of the most important developments in sequence analysis, integrated information retrieval, and molecular modeling that are harbingers of a new age of computational biology that is neither strictly experimental nor strictly theoretical, but may be more capable than either in dealing with the awesome complexity that living systems represent. Even at the present time, however, computational methods are of unquestionable value in the discovery of unanticipated connections among diverse biochemical systems and as rich sources of experimentally testable hypotheses about gene and protein structure and function. It will be interesting to see what another 6 years may bring.

I thank Dr. Manuel Peitsch for continuing collaboration and for his special insights into protein structure and biochemistry; Dr. Joseph Goldstein for helpful discussions and making sequence data available prior to publication; Dr. Jere Segrest for his encouragement; Drs. Steve Bryant, Jonathan Kans, and Greg Schuler for help with various computer matters; and Dr. Scott Grundy for his patience and for the opportunity to prepare this review.

Manuscript received 23 January 1992 and in revised form 10 March 1992.

REFERENCES

1. Boguski, M. S., M. Freeman, N. A. Elshourbagy, J. M. Taylor, and J. I. Gordon. 1986. On computer-assisted analysis of biological sequences: proline punctuation, consensus sequences, and apolipoprotein repeats. *J. Lipid Res.* **27:** 1011–1034.
2. Boguski, M. S., J. Ostell, and D. S. States. 1992. Protein and nucleic acid sequence databases and their uses. *In* Protein Engineering: A Practical Approach. A. R. Rees, R. Wetzel, and M. J. E. Sternberg, editors. IRL Press Ltd., Oxford, England. 57–88.
3. Doolittle, R. F., editor. 1990. Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences. *Methods Enzymol.* Vol. 183.
4. Gribskov, M., and J. Devereux, editors. 1991. Sequence Analysis Primer. W. H. Freeman, New York.
5. Adams, M. D., J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merril, A. Wu, B. Olde, R. F. Moreno, A. R. Kerlavage, W. R. McCombie, and J. C. Venter. 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science.* **252:** 1651–1656.
6. Adams, M. D., M. Dubnick, A. R. Kerlavage, R. Moreno, J. M. Kelley, T. R. Utterback, J. W. Nagle, C. Fields, and J. C. Venter. 1992. Sequence identification of 2,375 human brain genes. *Nature.* **355:** 632–634.
7. Waterston, R., C. Martin, M. Craxton, C. Huynh, A. Coulson, L. Hillier, R. Durbin, P. Green, R. Shownkeen, N. Halloran, T. Hawkins, R. Wilson, M. Berks, Z. Du, K. Thomas, J. Thierry-Mieg, and J. Sulston. 1992. A survey of expressed genes in *Caenorhabditis elegans*. *Nature Genet.* **1:** 114–123.
8. Benson, D., M. S. Boguski, D. J. Lipman, and J. Ostell. 1990. The National Center for Biotechnology Information. *Genomics.* **6:** 389–391.
9. Brown, M. L., C. Hesler, and A. R. Tall. 1990. Plasma en-

zymes and transfer proteins in cholesterol metabolism. *Curr. Opinion Lipidol.* **1:** 122–127.
10. Gray, P. W., G. Flaggs, S. R. Leong, R. J. Gumina, J. Weiss, C. E. Ooi, and P. Elsbach. 1989. Cloning of the cDNA of a human neutrophil bactericidal protein. *J. Biol. Chem.* **264:** 9505–9509.
11. Schumann, R. R., S. R. Leong, G. W. Flaggs, P. W. Gray, S. D. Wright, J. C. Mathison, P. S. Tobias, and R. J. Ulevitch. 1990. Structure and function of lipopolysaccharide binding protein. *Science.* **249:** 1429–1431.
12. Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.
13. Comer, D. 1988. Internetworking with TCP/IP: Principles, Protocols, and architecture. Prentice Hall, Englewood Cliffs, N.J.
14. Palca, J. 1990. Getting together bit by bit. *Science.* **248:** 160–162.
15. Maltese, W. A. 1990. Posttranslational modification of proteins by isoprenoids in mammalian cells. *FASEB J.* **4:** 3319–3326.
16. Rine, J., and S-H. Kim. 1990. A role for isoprenoid lipids in the localization and function of an oncoprotein. *New Biol.* **2:** 219–226.
17. Der, C. J., and A. D. Cox. 1991. Isoprenoid modification and plasma membrane association: critical factors for Ras oncogenicity. *Cancer Cells.* **3:** 331–340.
18. Glomset, J., M. Gelb, and C. Farnsworth. 1991. The prenylation of proteins. *Curr. Opinion Lipidol.* **2:** 118–124.
19. Reiss, Y., M. C. Seabra, S. A. Armstrong, C. A. Slaughter, J. L. Goldstein, and M. S. Brown. 1991. Nonidentical subunits of $p21^{H\text{-}ras}$ farnesyltransferase. *J. Biol. Chem.* **266:** 10672–10677.
20. Reiss, Y., J. L. Goldstein, M. C., Seabra, P. J. Casey, and M. S. Brown. 1990. Inhibition of purified $p21^{ras}$ farnesyl:protein transferase by Cys-AAX tetrapeptides. *Cell.* **62:** 81–88.
21. He, B., P. Chen, S-Y. Chen, K. L. Vancura, S. Michaelis, and S. Powers. 1991. RAM2, an essential gene of yeast, and RAM1 encode the two polypeptide components of the farnesyltransferase that prenylates a-factor and Ras proteins. *Proc. Natl. Acad. Sci. USA.* **88:** 11373–11377.
22. Goodman, L. E., C. M. Perou, A. Fujiyama, and F. Tamanoi. 1988. Structure and expression of yeast DPR1, a gene essential for the processing and intracellular localization of ras proteins. *Yeast.* **4:** 271–281.
23. Chen, W-J., D. A. Andres, J. L. Goldstein, D. W. Russell, and M. S. Brown. 1991. cDNA cloning and expression of the peptide-binding β subunit of rat $p21^{ras}$ farnesyltransferase, the counterpart of yeast DPR1/RAM1. *Cell.* **66:** 327–334.
24. Chen, W-J., D. A. Andres, J. L. Goldstein, and M. S. Brown. 1991. Cloning and expression of a cDNA encoding the α subunit of rat $p21^{ras}$ protein farnesyltransferase. *Proc. Natl. Acad. Sci. USA.* **88:** 11368–11372.
25. Kohl, N. E., R. E. Biehl, M. D. Schaber, E. Rands, D. D. Soderman, B. He, S. L. Moores, D. L. Pompliano, S. Ferro-Novick, S. Powers, K. A. Thomas, and J. Gibbs. 1991. Structural homology among mammalian and *Saccharomyces cerevisiae* isoprenyl-protein transferases. *J. Biol. Chem.* **266:** 18884–18888.
26. Ohya, Y., M. Goebl, L. E. Goodman, S. Peterson-Bjorn, J. D. Friesen, F. Tamanoi, and Y. Anraku. 1991. Yeast CAL1 is a structural and functional homologue to the DPR1 (RAM) gene involved in ras processing. *J. Biol. Chem.* **266:** 12356–12360.
27. Rossi, G., Y. Jiang, A. P. Newman, and S. Ferro-Novick. 1991. Dependence of Ypt1 and Sec4 membrane attachment on Bet2. *Nature.* **351:** 158–161.
28. Schuler, G. D., S. F. Altschul, and D. J. Lipman. 1991. A workbench for multiple alignment construction and analysis. *Proteins Struct. Funct. Genet.* **9:** 180–190.
29. Boguski, M. S., R. C. Hardison, S. Schwartz, and W. Miller. 1992. Analysis of conserved domains and sequence motifs in cellular regulatory proteins and locus control regions using new software tools for multiple alignment and visualization. *New Biol.* **4:** 247–260.
30. States, D. J., W. Gish, and S. F. Altschul. 1991. Improved sensitivity of nucleic acid database searches using application-specific scoring matrices. *Methods* (companion to *Methods Enzymol.*) **3:** 66–70.
31. States, D. J., and M. S. Boguski. 1991. Similarity and homology. Chapter 3. *In* Sequence Analysis Primer. M. Gribskov and J. Devereux, editors. W. H. Freeman, New York. 89–157.
32. Crawford, I. P., T. Niermann, and K. Kirschner. 1987. Prediction of secondary structure by evolutionary comparison: application to the alpha subunit of tryptophan synthase. *Proteins Struct Funct. Genet.* **2:** 118–129.
33. Ballester, R., D. Marchuk, M. Boguski, A. Saulino, R. Letcher, M. Wigler, and F. Collins. 1990. The NF1 locus encodes a protein functionally related to mammalian GAP and yeast IRA proteins. *Cell.* **63:** 851–859.
34. Lipman, D. J., S. F. Altschul, and J. D. Kececioglu. 1989. A tool for multiple sequence alignment. *Proc. Natl. Acad. Sci. USA.* **86:** 4412–4415.
35. Posfai, J., A. S. Bhagwat, G. Posfai, and R. J. Roberts. 1989. Predictive motifs derived from cytosine methyltransferases. *Nucleic Acids Res.* **17:** 2421–2435.
36. Hodgman, T. C. 1989. The elucidation of protein function by sequence motif analysis. *Comp. Appl. Biosci.* **5:** 1–13.
37. Locker, J., and G. Buzard. 1990. A dictionary of transcription control sequences. *DNA Sequence*-J. DNA Sequencing Mapping. **1:** 3–11.
38. Bairoch, A. 1991. PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Res.* **19:** 2241–2245.
39. Ghosh, D. 1990. A relational database of transcription factors. *Nucleic Acids Res.* **18:** 1749–1756.
40. Boguski, M. S., A. W. Murray, and S. Powers. 1992. Novel repetitive sequence motifs in the α and β subunits of prenyl-protein transferases and homology of the a subunit to the MAD2 gene product of yeast. *New Biol.* **4:** 408–411.
41. Reiss, Y., M. S. Brown, and J. L. Goldstein. 1992. Divalent cation and prenyl pyrophosphate specificities of the protein farnesyltransferase from rat brain, a zinc metalloenzyme. *J. Biol. Chem.* **267:** 6403–6408.
42. Luo, C-C., W-H. Li, M. N. Moore, and L. Chan. 1986. Structure and evolution of the apolipoprotein multigene family. *J. Mol. Biol.* **187:** 325–340.
43. Segrest, J. P., M. K. Jones, H. De Loof, C. G. Brouillette, Y. V. Venkatachalapathi, and G. M. Amantharamaiah. 1992. The amphipathic helix in the exchangeable apolipoproteins. *J. Lipid Res.* **33:** 141–166.
44. Schwartz, S., W. Miller, C-M. Yang, and R. C. Hardison. 1991. Software tools for analyzing pairwise alignments of long sequences. *Nucleic Acids Res.* **19:** 4663–4667.
45. Segrest, J. P., R. L. Jackson, J. D. Morrisett, and A. M. Gotto. 1974. A molecular theory of lipid-protein interactions in the plasma lipoproteins. *FEBS Lett.* **38:** 247–253.
46. Segrest, J. P., H. De Loof, J. G. Dohlman, C. G. Brouillette, and G. M. Ananthraramaiah. 1990. Amphipathic helix motif: classes and properties. *Proteins Struct. Funct. Genet.* **8:** 103–117.

47. Fasman, G. D. 1990. Protein conformational prediction. *In* Proteins: Form and Function. R. A. Bradshaw and M. Purton, editors. Elsevier Trends Journals, Cambridge. 135–145.
48. Luthy, R., and D. Eisenberg. 1990. Chapter 2. Protein. *In* Sequence Analysis Primer. M. Gribskov and J. Devereux, editors. W. H. Freeman, New York. 61–87.
49. Breiter, D. R., M. R. Kanost, M. M. Benning, G. Wesenberg, J. H. Law, M. A. Wells, I. Rayment, and H. M. Holden. 1991. Molecular structure of an apolipoprotein at 2.5-Å resolution. *Biochemistry.* **30:** 603–608.
50. Kanost, M. R., M. S. Boguski, M. Freeman, J. I. Gordon, G. R. Wyatt, and M. A. Wells. 1988. Primary structure of apolipophorin-III from the migratory locust, *Locusta migratoria:* potential amphipathic structures and molecular evolution of an insect apolipoprotein. *J. Biol. Chem.* **263:** 10568–10573.
51. Cornette, J. L., K. B. Cease, H. Margalit, J. L. Spouge, J. A. Berzofsky, and C. DeLisi. 1987. Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol.* **195:** 659–685.
52. Wilson, C., M. R. Wardell, K. H. Weisgraber, R. W. Mahley, and D. A. Agard. 1991. The three-dimensional structure of the LDL receptor-binding domain of human apolipoprotein E. *Science.* **252:** 1817–1822.
53. Boguski, M. S., N. A. Elshourbagy, J. M. Taylor, and J. I. Gordon. 1985. Comparative analysis of repeated sequences in rat apolipoproteins A-I, A-IV and E. *Proc. Natl. Acad. Sci. USA.* **82:** 992–996.
54. Boguski, M. S., N. A. Elshourbagy, J. M. Taylor, and J. I. Gordon. 1986. Rat apolipoprotein A-IV: application of computational methods for studying the structure, function and evolution of a protein. *Methods Enzymol.* **128:** 753–773.
55. Sali, A., J. P. Overington, M. S. Johnson, and T. L. Blundell. 1990. From comparison of protein sequences and structures to protein modelling and design. *In* Proteins: Form and Function. R. A. Bradshaw and M. Purton, editors. Elsevier Trends Journals, Cambridge. 163–171.
56. Blundell, T., D. Carney, S. Gardner, F. Hayes, B. Howlin, T. Hubbard, J. Overington, D. A. Singh, D. L. Sibanda, and M. Sutcliffe. 1988. 18th Sir Hans Krebs Lecture: Knowledge-based protein modelling and design. *Eur. J. Biochem.* **172:** 513–520.
57. Greer, J. 1990. Comparative modeling methods: application to the family of the mammalian serine proteases. *Proteins Struct. Funct. Genet.* **7:** 317–334.
58. Pearl, L. H., and W. R. Taylor. 1987. A structural model for the retroviral proteases. *Nature.* **329:** 351–354.
59. Wlodawer, A., M. Miller, M. Jaskolski, B. K. Sathyanarayana, E. Baldwin, I. T. Weber, L. M. Selk, L. Clawson, J. Schneider, and S. B. Kent. 1989. Conserved folding in retroviral proteases: crystal structure of a synthetic HIV-1 protease. *Science.* **245:** 616–621.
60. Sacchettini, J. C., J. I. Gordon, and L. J. Banaszak. 1988. The structure of crystalline *Escherichia coli*-derived rat intestinal fatty acid-binding protein at 2.5-Å resolution. *J. Biol. Chem.* **263:** 5815–5819.
61. Jones, T. A., T. Bergtors, J. Sedzik, and T. Unge. 1988. The three-dimensional structure of P2 myelin protein. *EMBO J.* **7:** 1597–1604.
62. Cowan, S. W., M. E. Newcomer, and T. A. Jones. 1990. Crystallographic refinement of human serum retinol binding protein at 2 Å resolution. *Proteins Struct. Funct. Genet.* **8:** 44–61.
63. Peitsch, M. C., and M. S. Boguski. 1991. The first lipocalin with enzymatic activity. *Trends Biochem. Sci.* **16:** 363.
64. Nagata, A. 1991. Human brain prostaglandin D synthase has been evolutionarily differentiated from lipophilic-ligand carrier proteins. *Proc. Natl. Acad. Sci. USA.* **88:** 4020–4024.
65. Keen, J. N., I. Caceres, E. E. Eliopoulos, P. F. Zagalsky, and J. B. C. Findlay. 1991. Complete sequence and model for the A2 subunit of the carotenoid pigment complex, crustacyanin. *Eur. J. Biochem.* **197:** 407–417.
66. Huber, R., M. Schneider, O. Epp, I. Mayr, A. Messerschmidt, J. Pflugrath, and H. Kayser. 1987a. Crystallization, crystal structure analysis and preliminary molecular model of the bilin binding protein from *Pieris brassicae*. *J. Mol. Biol.* **198:** 423–434.
67. Huber, R., M. Schneider, I. Mayr, R. Muller, R. Deutzman, F. Suter, H. Zuber, H. Falk, and H. Kayser. 1987b. Molecular structure of the bilin binding protein (BBP) from *Pieris brassicae* after refinement at 2.0 Å resolution. *J. Mol. Biol.* **198:** 499–513.
68. Newcomer, M. E., T. A. Jones, J. Aqvist, J. Sundelin, U. Eriksson, L. Rask, and P. A. Peterson. 1984. The three-dimensional structure of retinol-binding protein. *EMBO J.* **3:** 1451–1454.
69. Papiz, M. Z., L. Sawyer, E. E. Eliopolous, A. C. North, T. A. Jones, M. B. Newcomer, and P. J. Kraulis. 1986. The structure of beta-lactoglobulin and its similarity to plasma retinol binding protein. *Nature.* **324:** 383–385.
70. Peitsch, M. C., and M. S. Boguski. 1990. Is apolipoprotein D a mammalian bilin-binding protein? *New Biol.* **2:** 197–206.
71. Drayna, D., C. Fielding, J. McLean, B. Baer, G. Castro, E. Chen, L. Comstock, W. Henzel, W. Kohr, L. Rhee, K. Wion, and R. Lawn. 1986. Cloning and expression of human apolipoprotein D cDNA. *J. Biol. Chem.* **261:** 16535–16539.
72. McConathy, W. J., and P. Alaupovic. 1973. Isolation and partial characterization of apolipoprotein D: a new protein moiety of the human plasma lipoprotein system. *FEBS Lett.* **37:** 178–182.
73. Francone, O. L., A. Guraker, and C. Fielding. 1989. Distribution and functions of lecithin:cholesterol acyltransferase and cholesteryl ester transfer protein in plasma lipoproteins: evidence for a functional unit containing these activities together with apolipoproteins A-I and D that catalyzes the esterification and transfer of cell-derived cholesterol. *J. Biol. Chem.* **264:** 7066–7072.
74. Boyles, J. K., L. M. Notterpek, M. R. Wardell, and S. C. Rall. 1990. Identification, characterization, and tissue distribution of apolipoprotein D in the rat. *J. Lipid Res.* **31:** 2243–2256.
75. Pervaiz, S., and K. Brew. 1987. Homology and structure-function correlations between α_1-acid glycoprotein and serum retinol-binding protein and its relatives. *FASEB J.* **1:** 209–214.
76. Balbin, M., J. M. P. Freije, A. Fueyo, L. M. Sanchez, and C. Lopez-Otin. 1990. Apolipoprotein D is the major protein component in cyst fluid from women with human breast gross cystic disease. *Biochem. J.* **271:** 803–807.
77. Haagensen, D. E., G. Mazoujian, W. G. Dilley, C. E. Pedersen, S. J. Kister, and S. A. Wells. 1979. Breast gross cystic disease fluid analysis. I. Isolation and radioimmunoassay for a major component protein. *J. Natl. Cancer Inst.* **62:** 239–247.
78. Dilley, W. G., D. E. Haagensen, C. E. Cox, and S. A. Wells. 1990. Immunologic and steroid binding properties of the GCDFP-24 protein isolated from human breast gross cystic disease fluid. *Breast Cancer Res. Treat.* **16:** 253–260.

79. Lea, O. A. 1988. Binding properties of progesterone-binding cyst protein, PBCP. *Steroids.* **52:** 337–338.

80. Provost, P. R., P. K. Weech, N. M. Tremblay, Y. L. Marcel, and E. Rassart. 1990. Molecular characterization and differential mRNA tissue distribution of rabbit apolipoprotein D. *J. Lipid Res.* **31:** 2057–2065.

81. Chothia, C., and A. M. Lesk. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5:** 823–826.

82. Chothia, C., and A. M. Lesk. 1987. The evolution of protein structures. *Cold Spring Harbor Symp. Quant. Biol.* **52:** 399–405.

83. Lau, K. F., and K. A. Dill. 1990. Theory for protein mutability and biogenesis. *Proc. Natl. Acad. Sci. USA.* **87:** 638–642.