On the Classification and Evolution of Protein Modules

Hedvig Hegyi[1] and Peer Bork[1,2]

Received September 5, 1996

Our efforts to classify the functional units of many proteins, the modules, are reviewed. The data from the sequencing projects for various model organisms are extremely helpful in deducing the evolution of proteins and modules. For example, a dramatic increase of modular proteins can be observed from yeast to *C. elegans* in accordance with new protein functions that had to be introduced in multicellular organisms. Our sequence characterization of modules relies on sensitive similarity search algorithms and the collection of multiple sequence alignments for each module. To trace the evolution of modules and to further automate the classification, we have developed a sequence and a module alerting system that checks newly arriving sequence data for the presence of already classified modules. Using these systems, we were able to identify an unexpected similarity between extracellular C1Q modules with bacterial proteins.

KEY WORDS: Protein modules; sequence analysis; protein evolution; alerting system.

1. INTRODUCTION

With the sequencing of whole genomes a complete understanding of living cells has been envisaged; the pharmaceutical industry is already looking forward to new strategies of identifying drug targets resulting from the genome-based knowledge of complex cellular processes. However, the first step on the long road to understanding living cells remains the characterization of the function of individual gene products. This is primarily done by experimental work on the "proteome" in various organisms; molecular bioinformatics is supposed to support the characterization of particular gene products by predicting the function based on similarities to sequences stored in databases (Bork *et al.*, 1994*a*). Sequence databases indeed contain an enormous amount of valuable information that can be decoded, but as for today, current databases and automatic function prediction procedures are full of pitfalls (Bork and Bairoch, 1996) and much care has to be taken in functional predictions (see Bork, 1996, for an example).

Protein function is a loosely defined term, diverse functional properties such as molecular (e.g. "ATP-binding site"), cellular (e.g., "essential for secretion") and phenotypic features (e.g., "involved in wing development") are merged in sequence databases without a reliability index of these statements and often without quotation of the source. A further complication is the multifunctionality of many proteins. Metabolic enzymes with well-defined catalytic activities and detailed molecular knowledge comprise probably less than 25% of the human proteins (Ouzounis *et al.*, 1995). Most of the other proteins have regulatory or structural roles, are usually much longer and contain multiple domains with numerous distinct binding features. Often, each domain has an independent function and thus the identification and classification of individual domains should be most valuable for the understanding of the overall function of the numerous multidomain proteins. Here, we review our current understanding of the

[1] EMBL, 69012 Heidelberg, Germany and Max-Delbrück-Centrum for Molecular Medicine, Berlin-Buch, Germany.

[2] To whom correspondence should be addressed.

0277-8033/97/0700-0545$12.50/0 © 1997 Plenum Publishing Corporation

distribution of protein modules and their evolution; we also describe some software tools that are useful for the characterization and classification of modules and demonstrate this by the identification of C1q-like domains in bacteria.

2. DOMAINS AND MODULES

A *domain* is probably best defined as a spatially distinct structural unit that usually folds independently. In this definition, the sequence need not be contiguous. Thus, in the absence of structural knowledge, the term domain should not be used for regions that have been assigned a certain function, as the experimentally defined domain borders often do not coincide with the structurally determined ones (Bork *et al.*, 1996). Domains can already be found in ancient enzymes such as NADH-dependent dehydrogenases where the NADH-binding domain is merged with an independent substrate-binding domain. This appears to be the result of gene duplication, subsequent gene modification and gene fusion.

In more "modern" regulatory proteins that are not present in bacteria, archaebacteria or lower eukaryotes, several domains coexist within one protein and each domain can be found as "building block" in functionally diverse proteins. Such a spreading can no longer be simply explained by gene duplication and gene fusion, and thus the term *module* has been introduced. Modules are a subset of domains that are contiguous in sequence (Bork *et al.*, 1996). They are best characterized at the sequence level as they have identifiable amino acid patterns. The presence of an identified sequence region in another, otherwise unrelated protein and its location between other known modules are strong indicators for a module. Various modules are already known and their identification by sequence analysis often leads to a subsequent determination of their three-dimensional structure that, in turn, usually gives insights into their molecular interactions. "Exon shuffling" has been proposed as a major genetic mechanism leading to domain rearrangements within proteins (Gilbert, 1978), but with the completely sequenced genomes it becomes clear that modular proteins occur already in bacteria or organisms with no or only a few introns. Given the currently widely accepted "intron-late" theory, other genetic mechanisms for domain rearrangements are very likely to exist (Patthy, 1991, 1996). In any case, protein evolution by domain rearrangement is much more efficient than evolution by gene duplication and subsequent modification.

3. EVOLUTION OF MODULES AS INFERRED FROM THE ANALYSIS OF COMPLETE GENOMES

To understand the spread of the modules in the various regulatory and structural proteins, it is advisable to study their spread and their evolution. With the recent publication of complete genome sequences from all major phyla, this task has become much more feasible. If the complete gene pool of an organism such as baker's yeast is available and a module cannot be identified therein, it probably does not exist in this organism or the sequence similarity is too low to be able to detect it. In the latter case, a functional similarity is, however, rather unlikely. Large-scale sequence analysis of all putative gene products in an organism leads to functional classification of many of the proteins via sequence similarity and the distribution of functional classes has already been analyzed in several organisms (Fleischmann *et al.*, 1995; Fraser et al., 1995; Bult *et al.*, 1996; Dujon, 1996; Tatusov *et al.*, 1996). A shift in the contribution of certain functional classes of the complete gene pool can be observed during evolution: The fraction of metabolic enzymes gets smaller as the organisms become more complex; in contrast, the portion of regulatory proteins involved in communication increases (Ouzounis *et al.*, 1995, 1996). Modularity is clearly overrepresented in regulatory proteins. Furthermore, a recent study revealed that most of the positionally cloned disease genes contain modules and belong to the regulatory class (Mushegian *et al.*, 1997). As probably more than 40% of human proteins are involved in signal transduction and communication processes, the study of the structural basis of their multiple interaction is certainly necessary.

4. MODULES IN BACTERIAL PROTEINS

Sequence analysis of complete bacterial genomes has revealed that housekeeping enzymes such as DNA polymerases have already a modular architecture with up to four domains (Koonin and Bork, 1996). A considerable number of modules in bacteria have been reported, including the signal transducer and NifU families. The probably most abundant bacterial modular proteins are extracellular glycohydrolases, which contain in addition to their catalytic domains various carbohydrate-binding domains (Gilkes *et al.*; 1991). Horizontal gene transfer from animals (Little *et al.*, 1994) and recombination via hairpin structures in the DNA coding for linkers between domains (Wu *et al.*, 1990) are only two out of several proposed genetic mechanisms for domain spreading in prokaryotes.

5. CYTOPLASMIC MODULES THAT ARE WIDESPREAD IN YEAST

Although baker's yeast is a unicellular organism, many signal transduction mechanisms are already present and modules associated with these processes can be quantitatively traced, as the complete genome is available (Dujon, 1996). For example, more than 20 SH3 domains and several PH and WW domains are detectable in yeast. In addition to the various nuclear DNA-binding domains, there are now about 40 cytoplasmic domains known, many of them are already present in yeast. A considerable number of cytoplasmic proteins contain numerous successive repeats (e.g., ARM, HEAT, WD40, TPR, LRR, spectrin, ANK, etc.) that might have found their own way of evolution by slippage mechanisms at the DNA level. Several cytoplasmic domains that are abundant in animals have very few (e.g., SH2 domain) or no counterparts in yeast; for example, the PTB/PI domain (Bork and Margolis, 1995) that binds phosphotyrosine has not yet been identified in the yeast genome. As there is also no tyrosine kinase in yeast, regulation via phosphotyrosine has appeared probably later in the evolution.

6. EXTRACELLULAR MODULES

The most recent proteins are probably extracellular, as there was particular pressure with the emergence of multicellular organisms. Nearly 40% of human proteins might be extracellular or might have an extracellular part; most of them contain modules (Bork and Bairoch, 1995). The modularity of extracellular proteins had already been discovered in the early 1980s (for recent reviews see Baron *et al.*, 1991; Doolittle, 1995; Patthy, 1996; Bork *et al.*, 1996 and references therein). More than 60 extracellular modules have been classified so far (Bork and Bairoch, 1995; Bork *et al.*, 1996); many of them contain disulfide bridges and thus cannot be located in the reducing nuclear or cytoplasmic environment. Disulfide bridges not only stabilize the fold, but also allow a faster mutation rate, so that it is often very difficult to identify extracellular modules. As the remaining cysteine consensus pattern can also vary within a family (Bork *et al.*, 1994*a*; Bork *et al.*, 1996), methods more sensitive than standard database searches have to be applied to identify and classify modules.

7. TOWARD AUTOMATION OF MODULE CLASSIFICATION

The most alarming factor in sequence analysis is the increasing amount of data that flood the databases and that are often poorly annotated or even contain errors (Bork and Bairoch, 1996). In order to be able to keep track of the already identified domains and to update existing alignments, we have developed a module alerting system that scans weekly the newly incoming data to identify already known domains and to incorporate them into existing multiple alignments. As such an approach is equally useful for remaining informed on a particular single sequence, we first implemented a *sequence alerting system*, outlined in Fig. 1. It is an automated procedure which performs BLAST searches on the users' sequences against a daily protein database. This database is generated daily by comparing that day's nonredundant protein database with that of the previous day and by retaining only the new sequences; hence the name, Daily Differential Database. The nonredundant database is created using a script provided by the NCBI and contains only nonidentical proteins from the following databases: (1) SWISSPROT, (2) PDB, (3) TREMBL, (4) PIR, (5) GENPEPT, and (6) WORMPEPT. The server provides a BLASTP or a BLASTX search for the subscribed protein or DNA sequences, respectively. The users can subscribe their sequences via a WWW interface (URL: http://www.bork.embl-heidelberg.de/Alerting/). The users' requests consist of the sequences in FASTA format, and some parameters optionally determined by the users such as threshold for the score, filtering method, and the searching method (BLASTP or BLASTX). The users receive the results by e-mail. An e-mail is sent if the highest score in the result of the BLAST

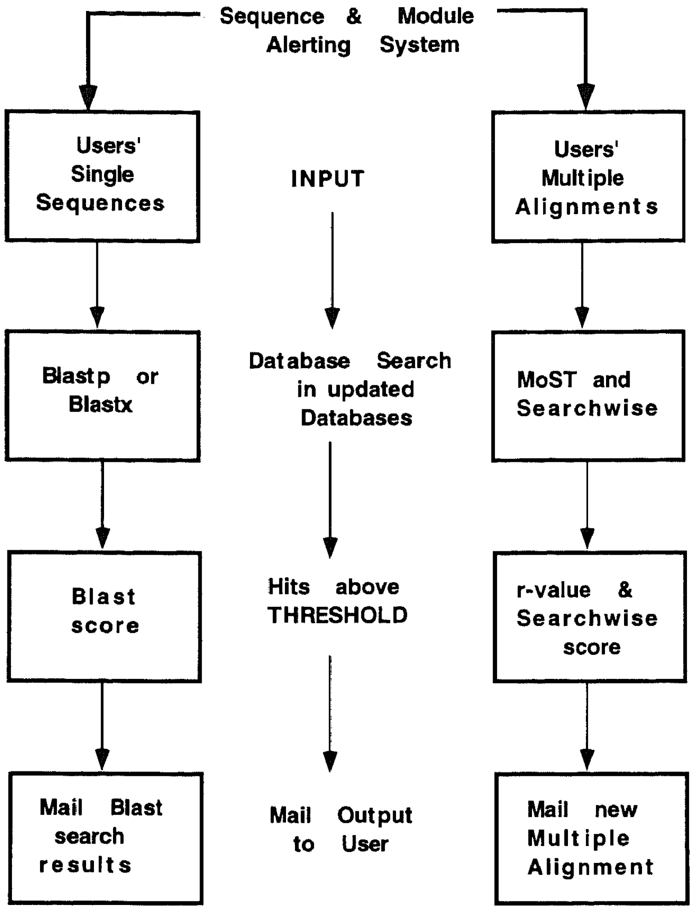
Sequence & Module Alerting System
Users' Single Sequences
INPUT
Users' Multiple Alignments
Blastp or Blastx
Database Search in updated Databases
MoST and Searchwise
Blast score
Hits above THRESHOLD
r-value & Searchwise score
Mail Blast search results
Mail Output to User
Mail new Multiple Alignment

Fig. 1. Flow chart of the Sequence and Module Alerting Systems. The two systems work independently, but have similar features. The Sequence Alerting System accepts a single sequence as an input via a WWW interface (URL: http://www.bork.embl-heidelberg.de/Alerting/). The system keeps the users' sequences for 1 year and checks for similar sequences every day in the Daily Differential Database (generated every day containing only the new database entries) using the Blastp (for proteins) or the Blastx (for DNAs) programs. The Module Alerting System requires a multiple alignment as input or offers subscription to several previously built alignments. MoST and SEARCHWISE programs are used to identify significant hits in the differential database.

search is above the user-defined threshold. The sequences are kept and followed for 1 year after subscription in our system unless the users delete them via the WWW interface. Users also can easily modify their subscribed sequences.

8. MODULE SEARCHING SYSTEM

In order to watch a whole domain family, we extended the alerting system to input a multiple alignment of a family of interest or to subscribe to an already characterized domain family. A semiautomatic procedure updates preexisting multiple alignments of module families, searching for new members of the families at regular intervals (weekly). To perform this, the system generates a Weekly Differential Database in the same fashion as described above for the daily database. The multiple alignment is the basis for iterative motif and profile searches (for details see Bork and Gibson, 1996); there is no recipe as for now to decide whether motif (here MoST [Tatusov *et al.*, 1994] is used) or profile searches (SEARCHWISE [Birney *et al.*, 1996] is implemented) will be more successful. The module alert offers both.

Motif searches concentrate on small conserved regions within the alignment to reduce the noise level (Bork and Gibson, 1996). The two most conserved blocks (ungapped regions) are defined in a multiple alignment and MoST searches in the weekly database are carried out. MoST (Tatusov *et al.*, 1994) is a motif-searching program which does an iterative search for all the short sequences, "motives" of the input block repeating the search until it does not find any new members in the database with the given, user-defined statistical significance (*r*-value, expected/observed ratio, was set as 0.05). The whole module-searching procedure is represented schematically in Fig. 1. Those database entries for which both regions of the multiple alignment give a hit with MoST are included automatically in the new multiple alignment, using the CLUSTALW program (Thompson *et al.*, 1994), the new motives found in the new entries are added to the input blocks, and the whole procedure is repeated with the extended input and the new Weekly Database every week. Those sequences for which only one of the two patterns was found have to be carefully examined to determine whether they are real members of the family or false positives.

Profile searches include all available information within an alignment to maximize the signal (Bork and Gibson, 1996). To confirm the MoST results and extend the search for all possible members of the family, a profile searching program, SEARCHWISE, is used. SEARCHWISE is part of the WiseTools package (Birney *et al.*, 1996); it performs a database search with a profile generated by PAIRWISE (another program of WiseTools package) from the multiple alignment of the module family. As SEARCHWISE does not calculate the statistical significance of the results, automatic runs currently rely on the thresholds provided by MoST.

```
                                                                   BLOCK 1                                                                                   BLOCK 2
                                         <------------------------------------->                                                                  <--------------->
HSCOLLX_1    605  GVTGMPVSAFTVILSKAYPA------IGTPIPFDKILYNRQQHYDPRTGIFTCQIPGIYYF--SYHVHVKGT--HVWVGLYKNGTP------VMYTYDEY--TKGYLDQASGSAIIDLTENDQVWLQLPNA-ESNGLYSSEYVHSSFSGFLVAPM------------ tr:X98568
MMCOL10A1_1  493  GVTGMPVSAFTVILSKAYPA------VGCPHPIYEILYNRQQHYDPRSGIFTCKIPGIYYF--SYHVHVKGT--HVWVGLYKNGTP------TMYTYDEY--SKGYLDTASGSAIMELTENDQVWLQLPNA-ESNGLYSSEYVHSSFSGFLVAPM------------ tr:X65121
S79214_1     29   GVTGMPVSAFTVILSKAYPA------VGCSHPIYEILYNRQQHYDPRSGIFTCKIPGIYYF--SYHIHVKGT--HVWVGLYKNGTP------TMYTYDEY--SKGYLGSASGSAIMELTENDQVWLQLPNA-ESNGLYSSEYVHSSFSGFLVA--------------- tr:S79214
C1QB_RAT     112  DYKATQKVAFSALRTVNSALR-----PNQAIRFEKVITNVNDNYEPRSGKFTCKVPGLYYF--TYHASSRG---NLCVNIVRGRDR------VMYTYDEY--IKGYLDQASGSAVIDLTENDQVWLQLPNA-GSNGLYSPEYVHSSFSGFLVAPM------------ sw:P23206
CA1A_BOVIN   539  GVTGMPVSAFTVILSKAYPA------IGTPIPFDKILYNKQQHYDPRTGIFTCKIPGIYYF--SYHIHVKGT--HAWVGLYKNGTP------VMYTYDEY--QKGYLDQASGSAVIDLMENDQVWLQLPNS-ESNGLYSSEYVHSSFSGFLFAQI------------ sw:P08125
CA1A_CHICK   539  ALTGMPVSAFTVILSKAYPG------ATVPIKFDKILYNRQQHYDPRTGIFTCRIPGLYYF--SYHVHAKGT--NVWVALYKNGSP------TMYTYDEY--SKGYLDQASGSAIMELTENDQVWLQLPNA-ESNGLYSSEYVHSSFSGFLVAPM------------ sw:Q05306
CA1A_MOUSE   544  GVTGMPVSAFTVILSKAYPA------VGAPIPFDEILYNRQQHYDPRSGIFTCKIPGIYYF--SYHVHVKGT--HVWVGLYKNGTP------VMYTYDEY--KKGFLDQASGSAVLLLRPGDRVFLQMPSE-QAAGLYAGQYVHSSFSGYLLYPM------------ sw:P14282
CA18_RABIT   609  GGPAYEMPAFTAELTAPFPP------VGAPIKFDRLLYNGRQNYNPQTGIFTCEVPGVYYF--AYHVHCKGG--NVWVALFKNNEP------VMYTYDEY--KKGFLDQASGSAVLLLRPGDRVFLQMPSE-QAAGLYAGQYVHSSFSGYLLYPM------------ sw:P27658
CA18_HUMAN   609  GGPAYEMPAFTAELTAPFPP------VGGPVKFNKLLYNGRQNYNPQTGIFTCEVPGVYYF--AYHVHCKGG--NVWVALFKNNEP------MMYTYDEY--KKGFLDQASGSAVLLLRPGDQVFLQNPFE-QAAGLYAGQYVHSSFSGYLLYPM------------ sw:Q00780
CA18_MOUSE   607  GGPAYEMPAFTAELTVPFPP------VGAPVKFDKLLYNGRQNYNPQTGIFTCEVPGVYYF--AYHVHCKGG--NVWVALFKNNEP------ATYTYDEY--KKGYLDQASGGAVLQLRPNDQVWVQIPSD-QANGLYSTEYIHSSFSGFLLCPT------------ sw:P25067
CA28_HUMAN   500  ELSAHATPAFTAVLTSPLPA------SGMPVKFDRTLYNGHSGYNPATGIFTCPVGGVYYF--AYHVHVKGT--NVWVALYKNNVP------ATYTYDEY--KKGYLDQASGGAVLQLRPNDQVWVQMPSD-QANGLYSTEYIHSSFSGFLLCPT------------ sw:P25318
CA28_MOUSE   35   ELSAHATPAFTAVLTSPFPA------SGMPVRFDRTLYNGHSGYNPATGIFSCPVGGVYYF--AYHVHVKGT--NVWVALYKNNVP--AQ-KVVTFCDYA--YN-TFQVTTGGMVLKLEQGENVFLQATDKNSLLGMEG---ANSIFSGFLLFPD------------ sw:P02746
C1QB_HUMAN   113  DYKATQKIAFSATRTINVPLR-----RDQTIRFDHVITNMNNNYEPRSGKFTCKVPGLYYF--TYHASSRG---NLCVNLMRGRERDSMQ-KVVTFCDYA--QN-TFQVTTGGVVLKLEQEEVVHLQATDKNSLLGIEG---ANSIFTGFLLFPD------------ sw:P14106
C1QB_MOUSE   113  DYRATQKVAFSALRTINSPLR-----PNQVIRFEKVITNANENYEPRNGKFTCKVPGLYYF--TYHASSRG---NLCVNLVRGRDRDRMQ-KVLTFCDYA--QN-TFQVTTGGVVLKLEQEEVVHLQATDKNSLLGVEG---ANSIFTGFLLFPDMDV---------- sw:P31721
C1QA_MOUSE   108  NIRDQPRPAFSAIRQNPMT-------LGNVVIFDKVLTNQESPYQNHTGRFICAVPGFYYF--NFQVISKW---DLCLFIKSSSGG-QPR-DSLSFSNTN--NKGLFQVLAGGTVLQLRRGDEVWIEK-DPAKGRIYQGT-EADSIFSGFLIFPS------------ sw:P98086
C1QA_HUMAN   108  NIKDQPRPAFSAIRRNPPM-------GGNVVIFDTVITNQEEPYQNHSGRFVCTVPGYYYF--TFQVLSQW---EICLSIVSSSRG-QVR-RSLGFCDTT--NKGLFQVVSGGMVLQLQQGDQVWVEK-DPKKGHIYQGS-EADSVFSGFLIFPS------------ sw:P02745
C1QC_HUMAN   112  RYKQKFQSVFTVTRQTHQPPA-----PNSLIRFNAVLTNPQGDYDTSTGKFTCKVPGLYYF--VYHASHTA---NLCVLLYRSGVK------VVTFCGHT--SK-TNQVNSGGVLLRLQVGEEVWLAVNDYYDMVGIQG---SDSVFSGFLLFPD------------ sw:P02747
C1QC_MOUSE   113  RYKQKHQSVFTVTRQTTQYPE-----ANALVRFNSVVTNPQGHYNPSTGKFTCEVPGLYYF--VYYTSHTA---NLCVHLNLNLAR------VASFCDHM--FN-SKQVSSGGALLRLQRGDEVWLSVNDYNGMVGIEG---SNSVFSGFLLFPD------------ sw:Q02105
CERB_HUMAN   55   VRSGSAKVAFSAIRSTNHEP-SEMSNRTMIIYFDQVLVNIGNNFDSERSTFIAPRKGIYSF--NFHVVKVYNRQTIQVSLMLNGWP------VISAFAGD--QDVTREAASNGVLIQMEKGDRAYLKLERG-NLMGGWK----YSTFSGFLVFPL------------ sw:P23435
CERL_RAT     85   VRSGSAKVAFSATRSTNHEP-SEMSNRTMTIYFDQVLVNIGNHFDLASSIFVAPRKGIYSF--SFHVVKVYNRQTIQVSLMQNGYP------VISAFAGD--QDVTREAASNGVLLLMEREDKVHLKLERG-NLMGGWK----YSTFSGFLVFPL------------ sw:P98087
HP20_TAMAS   64   KCPCRERSAFTVKFSGRLPP------PSEPVVFTEVLYNTQRDLKESTGVFNCVEPGNYHF--SFDVELYHC--KVKIGLMKNHIQ------VMEKHQLS--KNE-YENASGAMIMPLRQGDKVWLEADV--ETEEPDQA-KVVIYFSGFLISS------------- sw:Q06575
HP25_TAMAS   82   KCSSRPKSAFAVKLSERPPE------PFQPIVFKEALYNQEGHFNMATGEFSCVLPGVYNF--GFDIRLFQS--SVKIRLMRDGIQ------VREKEAQA--NDS-YKHAMGSVIMALGKGDKVWLESKL--KGTESEKG-ITHIVFFGYLLYGK------------ sw:Q06576
HP27_TAMAS   82   NCHSKGTSAFAVKANELPPA------PSQPVIFKEALHDAQGHFDLATGVFTCPVPGLYQF--GFHIEAVQR--AVKVSLMRNGTQ------VMEREAEA--QDG-YEHISGTAILQLGMEDRVWLENKL--SQTDLERG-TVQAVFSGFLIHEN------------ sw:Q06577
COLE_LEPMA   272  ESVEQIRSAFSVGLFSPRSFPP----PSLPVKFDKVFYNGEGHWDPTLNKFNVTYPGVYLF--SYHITVRNR--PVRAALVVNGVR------KLRTRDSL--YGQDIDQASNLALLHLTDGDQVWLETLR--DWNGXYSSSEDDSTFSGFLLYPDTKKPTAM----- sw:P98085
MM37222_1    108  EAAYMYRSAFSVGLETRVTV------PNVPIRFTKIFYNQQNHYDGSTGKFYCNIPGLYYF--SYHITVYMK--DVKVSLFKKDKA------VLFTYDQY--QEKNVDQASGSVLLHLEVGDQVWLQVYGDGDHNGLYADNVNDSTFTGFLLYHDTN---------- tr:U37222
HSUPST2_1    105  EGAYVYRSAFSVGLETYVTI------PNMPIRFTKIFYNQQNHYDGSTGKFHCNIPGLYYF--AYHITVYMK--DVKVSLFKKDKA------MLFTYDQY--QENNVDQASGSVLLHLEVGDQVWLQVYGEGERNGLYADNDNDSTFTGFLLYHDTN---------- tr:D45371
HS27109_1   1094  SYRYAPMVAFFASHTYGMTI------PG-PILFNNLDVNYGASYTPRTGKFRIPYLGVYVF--KYTIESFSA--HISGFLVVDGID-----KLAFESENINSEIHCDRVLTGDALLELNYGQEVWLRLAKGTIPAKFPP----VTTFSGYLLYRT------------ tr:U27109
YQCC_BACSU   204  WTPWQKISGFAHANIGTTGVQYLKKIDHTKIAFNRVIKDSHNAFDTKNNRFIAPNDGMYLIGASIYTLNYTSYINFHLKVYLNGKAYKTLHHVRGDFQEK--DNGMNLGLNGNATVPMNKGDYVEIWCYCNYGGDETLKR--AVDDKNGVFNFFDIQELGGRNYPRF sw:P45938
BSPBSXSE_25  525  WTAWQKISGFAHANIGTTGRQALIKGENNKIKYNRIIKDSHKLFDTKNNRFVASHAGMHLVSASLYIENTERYSNFELYVYVNGTKYKLMNQFRMPTPSNNSDNEFNATVTGSVTVPLDAGDYVEIYVYVGYSGD--VTR--YVTDSNGALNYFDVLELGGRNYPRV tr:Z70177
Consensus         ........sF.h..................h.h.th.hs.tt.hp.t.s.F.h...G.a.h...h.h........h.h.l..t..........h................s.hhh.h..tp.s.l........................Ghh...............
PHD vertebrate    LLL......EEEEE.LLLLLLLLLLLLL..EEEEEE..LLLLLLLLL.EEE.L..EEEE..EEEEEE.LL..EEEEEEEE.LLLLLLLL...EEE.............LL.EEEEE.LLLEEEEEELLLLLLLLL.LL........EEEEE.LLLLLLLL.....
PHD bacterial     ....HH....EEE..LL....EEEEE.LL..EEEEE.......LLLLL.EEE.LLL.EEEEEEEEEE......EEEEEEELL..E.......LLLLLL..LLLL ..EEEEEEE...LLL.EEEEEEE..LL.LLLL...EEE.L.L.EEEEE..H.LLLLLLLL
PHD all           LL.......EEEEE.LLLLLLLLLLLLL..EEEEEE..LLLLLLLLLLEEEE.L..EEE...EEEEEEE.L...EEEEEEE.LLLLLLLLL..EEE...LLLLLL....LL.EEEEE.LLLEEEEEEE.LLLLLLL..LL........EEEEE.LLLLLLLL.....
```

Fig. 2. Multiple alignment of the C1q family, constructed using CLUSTALW (Thompson *et al*., 1994). Only a very few regions were modified, suggested by the secondary structure predictions (Rost *et al*., 1994), conducted independently for the vertebrate and the bacterial sequences (last three sequences in the alignment). First and last columns show SWISSPROT (sw) or TREMBL (tr) identifiers and accession numbers, respectively. Database identifiers: C1QA, C1QB, C1QC, three different chains of C1Q; CA18, CA28, CA1A, hscollx, mmcoll0a1, S79214—1, collagens; CERB, cerebellin precursor; HP20—TAMAS, HP25—TAMAS, HP27—TAMAS, hybernation-specific proteins in chipmunk; COLE—LEPMA, inner-ear-specific collagen from bluegill; mm37222—1, complement-related protein Acrp30; hsupst2—1, collagen-like factor apM1; YQCC—BACSU, BSPBSXSE—25, hypothetical proteins in "skin element" region in *Bacillus subtilis*. Invariant residues are shown in bold. The consensus line immediately beneath the alignment indicates residues or amino acid properties conserved: h, hydrophobic; l, aliphatic; p, polar; s, small; u, tiny; t, turn-like (polar or tiny). PHD secondary structure predictions H (helix), E (strand), L (loop) with >80% expected accuracy (Rost *et al*., 1994) are shown for the two subgroups (vertebrate and bacteria) and for all (last line).

9. AN EXAMPLE: IDENTIFICATION OF THE C1q DOMAIN IN BACTERIAL PROTEINS

To give an example that shows how useful such a module-alerting system can be, we describe here the identification of bacterial homologues of the C1q domain. As extracellular domains have so far not been found in unicellular organisms, the presence of C1q domains in bacteria is quite surprising. C1q is a subunit of the C1 enzyme complex and plays a key role in the recognition of immune complexes, initiating the classical pathway of complement activation. The molecule consists of an N-terminal nonhelical region, a triple-helical (collagenous) region and a C-terminal · globular head (as shown by electron microscopy studies), which is called the C1q module. It was also concluded from Fourier-transform infrared spectroscopy studies that the C1q module contains primarily β-sheet structure (Smith *et al.*, 1994). According to these authors, the module consists of about 136 amino acids, which probably form ten β strands interspersed by β-turns and/or loops. In addition to the complement component C1q, the module was found in several different proteins of animals (Fig. 2) such as vertebrate collagen type VIII and X, chipmunk hibernation-associated plasma proteins HP-20, HP-25, and HP-27, as well as human multimerin, a large soluble protein found in platelets and in the endothelium of blood vessels (Hayward *et al.*, 1995). C1q modules appear to be located exclusively at the C-terminal end of the respective proteins (Fig. 3).

In Fig. 2, the two conserved blocks are indicated that were used for the searches of new members of the C1q family. Based on a training set of five members, all but three sequences were automatically aligned; only multimerin (Hs27109—1) and the two bacterial sequences (YQCC—BACSU and BSPBSXSE—25) were aligned separately, as they only matched the first pattern. Multimerin has already been identified to contain a C1q domain (Hayward *et al.*, 1995); the last two sequences are hypothetical proteins from *Bacillus subtilis*. They are located in the so-called 'skin element' (skin = s̲igK̲ in̲tervening element), a 42 kb DNA which contains no genes essential for viability (Takemaru *et al.*, 1995).

In all of the last three sequences the second pattern is not strong enough to be detected by MoST, but HS27109—1 and YQCC—BACSU were also found by the SEARCHWISE program. Although the similarity between the bacterial and

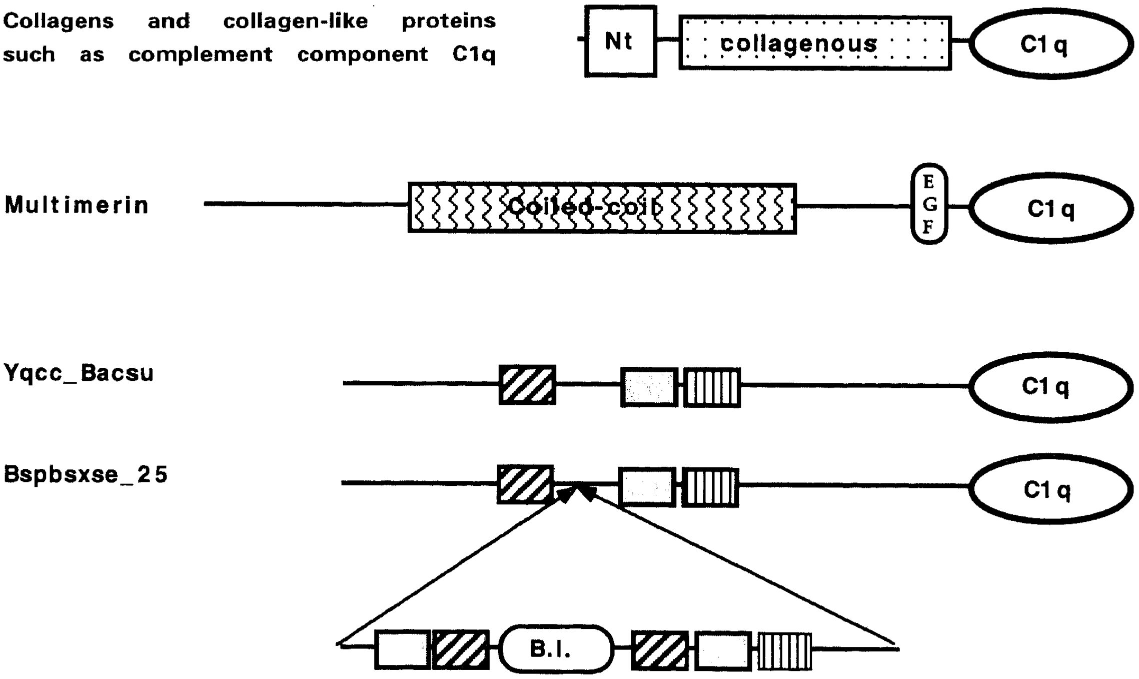

Fig. 3. Example C1q domain: modular architecture in diverse proteins. The hatched and gray boxes indicate recurring short patterns, which are also present in the region inserted in the *B. subtilis* protein BSPBSXSE—25. Apart from the insert, the two *B. subtilis* proteins have an average of 66% identity. The insert contains a region (marked B.I.) which has sequence similarity with a *Bacillus licheniformis* protein (id: BLD712).

vertebrate sequences is too low to be detected by conventional sequence searching methods like BLAST, it is clear from the multiple alignment (Fig. 2) that the bacterial sequences fit the overall C1q domain structure. Our findings are supported by secondary structure analysis, using the PhD server (Rost *et al.*, 1994). It predicts all ten β-strands proposed by Smith *et al.* (1994). Smith *et al.* (1994) not only used several different secondary structure prediction methods, but also gained experimental evidence (by infrared spectroscopy) for the β-sheet structure of the C1q domain.

10. CONCLUSION

Comparative sequence analysis is a powerful methodology to identify, classify, and trace the history of protein modules. It is greatly facilitated by the sequencing of complete genomes of model organisms and with an ongoing automation of the sequence analysis. As an example, we have introduced a module updating system that allows one to keep track of a particular family of interest. Unfortunately, specific biological functions cannot be directly associated with the identification of a protein module. Modules might have different functions in different settings, there is often not enough functional information available, and modules sometimes just do not function autonomously. Nevertheless, a classification of existing modules is an essential prerequisite to the understanding of the overall function of multidomain proteins.

REFERENCES

Baron, M., Norman, D. G., and Campbell, I. D. (1991). *TIBS* **16,** 13–17.

Birney, E. Thompson, J., and Gibson, T. (1996). *Nucleic Acids Res.* **24,** 2730–2739.

Bork, P., and Bairoch, A. (1995). *TIBS* **20**(3) (Poster Supplement C02).

Bork, P., and Bairoch, A. (1996). *Trends Genet.* **12,** 425–427.

Bork, P., and Gibson, T. (1996). *Meth. Enzymol.* **266,** 161–183.

Bork, P. (1996). *Science* **271,** 1431–1432.

Bork, P., and Margolis, B. (1995). *Cell* **80,** 693–694.

Bork, P., Holm, L., and Sander, C. (1994*a*). *J. Mol. Biol.* **242,** 309–320.

Bork, P., Ouzounis, C., and Sander, C. (1994*b*). *Curr. Opin. Struct. Biol.* **4,** 393–403.

Bork, P., Downing, K. A., Kieffer, B., and Campbell, I. D. (1996). *Quart. Rev. Biophys.* **29,** 119–167.

Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleiscmann, R. D., Sutton, G. G., Blake, J. A., Fitzgerald, L. M., Clayton, R. A., and Gocayne, J. D. (1996). *Science* **273,** 1058–1073.

Doolittle, R. F. (1995). *Annu. Rev. Biochem.* **64,** 287–314.

Dujon, B. (1996). *Trends Genet.* **12,** 263–270.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M. *et al.* (1995). *Science* **269,** 496–512.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M. *et al.* (1995). *Science* **270,** 397–403.

Gilbert, W. (1978). *Nature* **271,** 501.

Gilkes, N. R., Henrissat, B., Kilburn, D. G., Miller, R. C. Jr., and Warren, R. A. (1991). *Microbiol. Rev.* **55,** 303–315.

Hayward, C. P. M., Hassell, J. A., Denomme, G. A., Rachubinski, R. A., Brown, C., and Kelton, J. G. (1995). *J. Biol. Chem.* **270,** 18246–18251.

Koonin, E. V., and Bork, P. (1996). *TIBS* **21,** 128–129.

Little, E., Bork, P., and Doolittle, R. F. (1994). *J. Mol. Evol.* **39,** 631–643.

Mushegian, A. R, Bassett, D. E. Jr., Boguski, M. S., Bork, P., and Koonin, E. V. (1997). *Proc. Natl. Acad. Sci.* **94,** in press.

Ouzounis, C., Valencia, A., Tamames, J., Bork, P., and Sander, C. (1995). In *Advances in Artificial Life* (Moran, F., Moreno, A., Merelo, J. J., and Chacon, P. eds.), Springer-Verlag, Berlin, pp. 843–852.

Ouzounis, C., Casari, G., Sander, C., Tamames, J., and Valencia, A. (1996). *TIBTECH* **14,** 280–285.

Patthy, L. (1991). *Bioessays* **13,** 187–192.

Patthy, L. (1996). *Matrix Biol.* **15,** 301–310.

Rost, B., Sander, C., and Schneider, R. (1994). *CABIOS* **10,** 53–60.

Smith, K. F., Haris, P. I., Chapman,,D., Reid, K. B. M., and Perkins, S. J. (1994). *Biochem. J.* **301,** 249–256.

Takemaru, K., Mizuno, M., Sato, T., Takeuchi, M., and Kobayashi, Y. (1995) *Microbiology* **141,** 323–327.

Tatusov, R. L., Altschul, S. F., and Koonin, E. V. (1994). *Proc. Natl.Acad. Sci. USA* **91,** 12091–12095.

Tatusov, R. L., Mushegian, A. R., Bork, P., Brown N. P., Hayes, W. S., Borodovsky, M., Rudd, K. E., and Koonin, E. V. (1996). *Curr. Biol.* **6,** 279–291.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) *Nucleic Acids Res.* **22,** 4673

Wu, L. F., Tomich, J. M., and Saier, M. H. Jr. (1990) *J. Mol. Biol.* **213,** 687–703.