Protein sequence motifs

Peer Bork* and Eugene V Koonin†

Protein sequence motifs are signatures of protein families and can often be used as tools for the prediction of protein function. The generalization and modification of already known motifs are becoming major trends in the literature, even though new motifs are still being discovered at an approximately linear rate. The emphasis of motif analysis appears to be shifting from metabolic enzymes, in which motifs are associated with catalytic functions and thus often readily recognizable, to structural and regulatory proteins, which contain more divergent motifs. The consideration of structural information increasingly contributes to the identification of motifs and their sensitivity. Genome sequencing provides the basis for a systematic analysis of all motifs that are present in a particular organism. A systematically derived motif database is therefore feasible, allowing the classification of the majority of the newly appearing protein sequences into known families.

Addresses
*European Molecular Biology Laboratory, Meyerhofstrasse 1, 69012 Heidelberg, and Max-Delbrück Center, 13189 Berlin-Buch, Germany; e-mail: bork@embl-heidelberg.de
†National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Current Opinion in Structural Biology 1996, **6**:366–376

Abbreviations

BIR	baculovirus IAP repeat
FAD	flavin adenine dinucleotide
IAP	inhibitor of apoptosis protein
PKD	polycystic kidney disease
PPI	peptidylprolyl *cis–trans* isomerase
SH	Src homology
TNF	tumor necrosis factor

Introduction

With the exponential growth of the amount of sequenced DNA and consequently of the identified gene products, the sequence-based classification of all proteins becomes a major issue. One of the most successful approaches is to define signatures of protein families that unambiguously retrieve all the members of the respective family from the complete sequence database and allow the classification of new proteins into these families. The signatures can be derived as simple consensus patterns, or more complex descriptors, such as profiles or blocks (for clarification of these and other terms related to similarity searches see [1•]). Searches with conserved motifs aim at the reduction of the higher noise level from the more variable regions of the alignment. An alternative strategy is to maximize the overall signal of a family as implemented in profile searches that include each position of a multiple sequence alignment [1•,2]. Both the motif and the profile approaches have been successfully applied to protein identification and classification. Database searches with short motifs are more amenable to statistical analysis and usually much faster than profile searches. Therefore, motifs are frequently used as signatures for protein families in protein/domain databases such as PROSITE, BLOCKS, PRINTS, SBASE [3•,4–6] and PFAM (E Sonnhammer, personal communication).

What is a protein sequence motif?

The term motif applied to sequence analysis is rather vague and only implies the conservation of short regions within larger sequences. Thus, there are different meanings and interpretations attached to it.

Firstly, there are short functional motifs that only consist of a very small number of specific residues and that have mostly evolved independently from the surrounding structural context (e.g. myristilation sites, glycosylation sites, Src homology [SH]2-binding sites).

Secondly, there are short structural motifs that reflect certain topological constraints at the sequence level (e.g. N and C caps of α helices), but that are often not specific enough to be routinely used in structural predictions and in sequence analysis.

Thirdly, there are functional motifs that do not rely on invariant residues but are somewhat more constrained at the sequence level, such as transmembrane regions, signal sequences, and cell sorting or other recognition signals.

Finally, the majority of the motifs discussed in the literature are unique, detectable sequence features that distinguish a specific set of protein sequences from the rest of the protein universe. Such motifs reflect functional and structural constraints and imply a common descent (homology) for the given group of proteins or domains. A few of these motifs have been identified experimentally but, mostly, it is the sequence analysis that has led to the delineation of such motifs.

The first three motif types are mostly independent from the concept of homology and are not further discussed here.

Conserved motifs can already be defined on the basis of alignments of orthologs, that is, proteins with identical functions in different species. Such motifs are useful for the prediction of key functional residues, especially when orthologous sequences from a phylogenetically diverse set of organisms are available. The examples discussed below, however, describe conserved regions in divergent sequences with different overall functions, that is, they include a number of paralogous proteins [7]. In some cases, these sequences are so divergent that one can no longer be confident that the motif conservation reflects common ancestry; rather, it may be the result of convergent evolution towards similar binding properties [8].

In addition to motifs that unite very divergent protein superfamilies, each family frequently contains unique motifs (e.g. for a particular substrate-binding site) that distinguish it from all the other families. This kind of motif will attract more interest when all the motifs defining large protein superfamilies are identified.

The term 'motif' is not used by all authors describing relationships between protein sequences. Key terms such as 'family', 'subfamily', and 'superfamily' also have different meanings in the literature, and at present, we cannot give them robust definitions. How do we include knowledge on domains into the family concept, that is, independent folding units within larger proteins? Where does a sequence family end, and where does the next one begin? When should we cluster some well-defined families together to form a superfamily that still contains common, albeit more degenerate, motifs? The last point becomes more and more problematic with the wealth of data on three-dimensional (3D) structures of proteins, the comparison of which may suggest a common ancestry between proteins or domains that is not always mirrored at the sequence level (for review see [9]).

What is a 'new' protein sequence motif?

Strictly speaking, a new motif is derived when several proteins are grouped together for the first time by similarity searches, and it is shown that they share at least one conserved motif (signature) that is stringent enough to retrieve all or most of the family members from the complete protein sequence database. On the other hand, the motif should not be defined too rigidly, so as to allow the detection of new members of a protein family distantly related to the original ones. To a large extent, the definition of a 'new' motif is a matter of semantics, because many important motifs that have been described recently are significant variations on previously identified ones. The two most common sequence motifs in proteins are the nucleotide-binding P-loops and the nucleic acid-binding zinc fingers; both of these motifs exist in numerous variants, and the identification of another one is sometimes an important discovery.

This illustrates another problem: what shall we call a variant motif and what constitutes a separate motif? For the same protein family, the answers may depend on the motif-search method. The example in Figure 1 raises several other questions regarding the identification of new motifs. If a motif is already known for a family, shall we call a second one in the same family 'novel'? If a motif has been described for a very small subset of a family, does its generalization or modification justify the term 'novel'? Figure 1 shows a second motif in many of the GAL4-like transcription factors that have been previously grouped together because of their common DNA-binding domain (for a recent review see [1•]); the 'new' motif might contain the dimerization site.

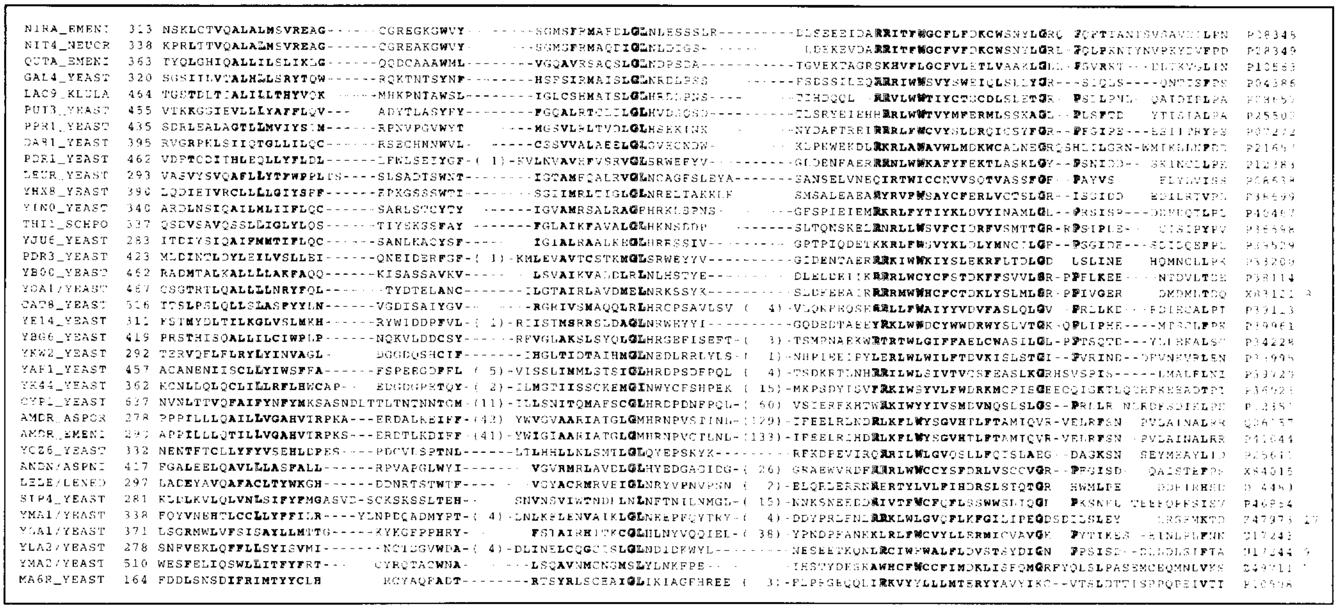
The recent literature includes numerous examples of new motifs, the identification of new, deviant versions of already known motifs, and the generalization of known motifs to group together protein families previously considered to be unrelated. Many reports concentrate only on adding a single new member to an existing family as new, potentially important functional and structural insights emerge for this particular sequence or the whole family. Here we focus primarily on reports of new motifs but also include a few examples of significant extensions, generalizations, and functional re-interpretations of known motifs.

Despite all the problems in defining a 'novel' motif, it seems to be possible to develop a minimal set of criteria to judge claims of motif discoveries [10•]; more rigorous methods for their detection should simplify this task.

Methods for the identification of sequence motifs

The methods used vary from identification 'by eye' to automatic delineation from an initial standard database search with one query protein using programs like Blastp [11]. Several methodological aspects of motif and profile analysis have recently been summarized [1•]. There are two extreme descriptions of motifs: namely, strings and matrices. String analysis is the simplest approach, under which motifs are defined as amino acid residues separated by fixed spacers. Such simple patterns may become more complex in that multiple residues may match a particular position and that variable spacers are allowed. The PROSITE library is the most popular and, to our knowledge, the best-annotated pattern collection that recently also extended the motif description complexity [3•]. The amino acid (single letter code) pattern for the P-loop, (AG)×4GK(ST) is a classical example of a simple motif (string) description whereby (AG) means amino acid A or G and ×4 denotes the length of the spacer. Patterns can be more flexible in that weights are assigned to particular positions, and spacers of variable length are allowed. Methods along these lines have been reported recently [11–16]. Matrix analysis involves a transformation of an (ungapped) alignment block into a position-dependent weight matrix which is then used to screen the database (for details, see [1•,17]). The outcome of the search strongly depends on the method used to construct the matrix. The methods that attempt to extrapolate the amino acid frequencies in the given block to approximate the distribution in the entire family have a clear advantage over the averaging and weighting using amino acid substitution tables [17].

Figure 1



```
NIRA_EMENI  313 NSKLCTVQALALMSVREAG-----CGREGKGWVY------SGMSFPMAFDLGLNLESSSLR---------DLSEEEIDARRITFWGCFLFDKCWSNYLGRQ FQFTIANISVSAVDILPN  P28345
NIT4_NEUCR  338 KPRLTTVQALALMSVREAG------CGREAKGWVY    ---SGMGFRMAQTIGLNLDIGS- -----    LDEKEVDARRITFWGCFVFDKCWSNYLGRL-FQLPKNIYNVPKYDVFPD  P28349
QUTA_EMENI  363 TYQLGHIQALLILSLIKLG-----QQDCAAAWML--------VGQAVRSAQSLGLNDPSDA-----------TGVEKTAGRSKHVFLGCFVLETLVAAKLGLL-FGVRKT---DLTKVGLIN  P10563
GAL4_YEAST  320 SGSIILVTALHLLSRYTQW------RQKTNTSYNF--------HSFSIRMAISLGLNRDLPSS   ------FSDSSILEQRRRIWWSVYSWEIQLSLLYGR---SIQLS----QNTISFPS  P04386
LAC9_KLULA  464 TGSTDLTIALILLTHYVQK -----MHKPNTAWSL--------IGLCSHMATSLGLHRDLPNS----   --TIHDQQL RRVLWWTIYCTDCDLSLETGR PSILPNL  QATDIPLPA  P08657
PUT3_YEAST  455 VTKKGGIEVLLLYAFFLQV---    ADYTLASYFY--------FGQALRTCLILGLHVDSQSD-----------TLSRYEIEHHRRLWWTVYMFERMLSSKAGL--PLSFTD    YTISIALPA  P25502
PPR1_YEAST  435 SDRLEALAGTLLMVIYSIM------RPNVPGVWYT     ----MGSVLRLTVDLGLHSEKINK     --NYDAFTREIRRRLFWCVYSLDRQICSYFGR--PFGIPE---ESIIPRYPS  P07272
DA81_YEAST  395 RVGRPKLSIIQTGLLILQC------RSECHNNWVL-  -----CSSVVALAEELGLQVECNDW-    -- KLPKWEKDLRKRLAWAVWLMDKWCALNEGRQSHLILGRN-WMIKLLNFDD  P21657
PDR1_YEAST  462 VDFTCDITHLEQLLYFLDL------LFKLSEIYGF-( 1)-KVLNVAVHFVSRVGLSRWEFYV----------GLDENFAERRRNLWWKAFYFEKTLASKLGY--PSNIDD-- SKINCLLPK  P12383
LEU3_YEAST  293 VASVYSVQAFLLYTFWPPLTS----SLSADTSWNT---------IGTAMFQALRVGLNCAGFSLEYA-------SANSELVNEQIRTWICCNVVSQTVASSFGF--PAYVS     FLYLVISS  P08638
YHX8_YEAST  390 LQDIETVRCLLLLGIYSFF------EPKGSSSWTI---------SGIIMRLTIGLGLNRELTAKKLF      SMSALEAEARYRVPWSAYCFERLVCTSLGR--ISGIDD   EDILRTVPL  P38699
YIN0_YEAST  340 ARDLNSIQAILMLIIFLQC------SARLSTCYTY  ------IGVAMRSALRAGFHRKLSPNS-    ---GFSPIEIEMRKRLFYTIYKLDVYINAMLGL--PRSISP---DEPEQTLPL  P40467
THI1_SCHPO  337 QSDVSAVQSSLLIGLYLQS------TIYSKGSFAY -------FGLAIKFAVALGLHKNSDDP  -------SLTQNSKELRNRLLWSVFCIDRFVSMTTGR-RPSIPLE--  CISIPYPV  P36598
YJU6_YEAST  283 ITDIYSIQAIFMMTIFLQC------SANLKACYSF----     IGIALRAALKEGLHRESSIV--------- GPTPIQDETKKRLFWSVYKLDLYMNCILGF--PSGIDE---SDIDQEFPL  P39529
PDR3_YEAST  423 MLDINTLDYLEILVSLLEI------QNEIDERFGF-( 1)-KMLEVAVTCSTKMGLSRWEYYV----------GIDENTAERRRKIWWKIYSLEKRFLTDLGD  LSLINE    HQMNCLLPK  P33200
YB0C_YEAST  462 RADMTALKALLLLAKFAQQ------KISASSAVKV-----    LSVAIKVADDLRLNLHSTYE--------  DLELDEIIKRRRLWCYCFSTDKFFSVVLGR-PPFLKEE---NTDVLTDE  P38114
YOA1/YEAST  467 CSGTRTLQALLLLNRYFQL-------TYDTELANC-------ILGTAIRLAVDMELNRKSSYK---------SLDFEEAIRRRRMWWHCFCTDKLYSLMLGR-PPIVGER     DMDMLTDQ  X83121
CAT8_YEAST  516 TTSLPSLQLLSLASFYYLN------VGDISAIYGV--   ----RGRIVSMAQQLRLHRCPSAVLSV (  4)-VLQKFHQSERRLLFWAIYYVDVFASLQLGV  PRLLKD---FDIECALPT  P39113
YE14_YEAST  311 FSTMYDLTILKGLVSLMKH------RYWIDDPFVL-( 1)-RIISTMSRRSLDAGLNRWEYYI----------GQDEDTAEEYRKLWWDCYWWDRWYSLVTGK-QPLIPHE----MPSGLFPK  P39961
YBG6_YEAST  419 PRSTHISQALLILCIWPLP------NQKVLDDCSY------RFVGLAKSLSYQLGLHRGEFISEFT-(  3)-TSMPNAEKWRTRTWLGIFFAELCWASILGL-PPTSQTD---YLIHKALSC  P34228
YKW2_YEAST  292 TERVDFLFLRYLYINVAGL-      DGGDQSHCIF-------IHGLTIDTAIHMGLNEDLRRLYLS-(  1)-NHPIEEIPYLERLWLWILFTDVKISLSTGI--PVRIND--DFVNEVPLSN  P35995
YAF1_YEAST  457 ACANENIISCLLYIWSFFA------FSPEEGDFFL ( 5)-VISSLIMMLSTSIGLHRDPSDFPQL (  4)-TSDKRTLNHRRILWLSIVTVCSFEASLKGRHSVSPIS-----LMALFLNI  P39720
YK44_YEAST  362 KCNLLQLQCLILLRFLHWCAP--   EDGDGPETQY-( 2)-ILMGTIISSCKEMGINWYCFSHPEK ( 15)-MKPSDYIGVFRKIWSYVLFWDRKMCFISGEECQIGKTLQIHPKEEADTPI  P36023
CYP1_YEAST  637 NVNLTTVQFAIFYNFYMKSASNDLTTLTNTNNTGM-(11)-ILLSNITQMAFSCGLHRDPDNFPQL-( 60) VSIERFKHTWRKIWYYIVSMDVNQSLSLGS--PRLLR NLRDFSDIKLPS  P12351
AMDR_ASPOR  278 PPPILLLQAILLVGAHVIRPKA---ERDALKEIFF-(42) YWVGVAARIATGLGMHRNPVSTINL-(129)-IFEELRLNDRLKFLWYSGVHTLFTAMIQVR-VELRFSN  PVLAINALRR  Q06157
AMDR_EMENI  293 APPILLLQTILLVGAHVIRPKS---ERDTLKDIFF-(41)-YWIGIAARIATGLGMHRNPVCTLNL-(133)-IFEELRIHDRLKFLWYSGVHTLFTAMIQVR-VELRFSN  PVLAINALRR  P41044
YCZ6_YEAST  332 NENTFTCLLYFYVSEHLLDPES---PDCVLSPTNL------LTLHHLLKLSMTLGLQYEPSKYK---------RFKDPEVIRQRRILWLGVQSLLFQISLAEG--DAGKSN  SEYMEAYLID  P25611
ANDN/ASPNI  417 FGALEELQAVLLLASFALL------RPVAPGLWYI-------  VGVRMRLAVDLGLHYEDGAGIDG-( 26) GRAEWVRDFRRRLWWCCYSFDRLVSCCVGR--PFGISD   QAISTEFPS  X84015
LELE/LENED  297 LADEYAVQAFACLTYWKGH------DDNRTSTWTF---  -  VGYACRMRVEIGLNRYVPNVPSN (  2)-ELQRLERRNRERTYLVLPIHDRSLSTQTGR  HWMLPE    DDFIRHSD  D44480
SIP4_YEAST  281 KLDLKVLQLVNLSIFYPMGASVD-SCKSKSSLTEH--     SNVNSVIWTNDILNLNFTNILNMGL-( 15)-NNKSNEEDDRIVTFWCFQFLSSWWSLIQGI  PKSNEL TEEFQPFSISV  P46954
YMA1/YEAST  338 FQYVNEHTLCCLLYFFILR----YLNPDQADMYPT-( 4)-DNLKFLENVAIKLGLNEEPFQYTRY-(  4)-DDYPRLFNLRRKLWLGVQFLKFGILIPEGDSDILSLEY     LRSPMKTD  Z47973 17
YLA1/YEAST  371 LSGRMWLVFSISAYLLMTTG-----KYKGFPPHRY-      FSIAIRHITKCGLHLNYVQQIEL-( 38)-YPNDPFANKKLRLFWCVYLLERMICVAVGK  PYTIKES -EINLPLFNK  U17243
YLA2/YEAST  278 SNFVEKLQFFLLSYISVMI-------NCIDGVWDA-( 4)-DLINELCQGCISLGLNDIDKWYL---------NESEETKQNLRCIWFWALFLDVSTSYDIGN  PPSISD--DLDDLSIPTA  U17244 9
YMA2/YEAST  510 WESFELIQSWLLITFYFRT---    CYRQTACWNA-    ---LSQAVNMCNGMSLYLNKFPE--       -IHSTYDESKAWHCFWCCFIMDKLISFQMGRFYQLSLPASEMCEQMNLVKS  Z49211
MA6R_YEAST  164 FDDLSNSDIFRIMTYYCLH        RCYAQFADT--------RTSYRLSCEAIGLIKIAGFHREE (  3)-FLPEGEQQLIRKVYYLLLMTERYYAVYIKC--VTSLDTTISPPQPEIVTI  P10508
```

The alignment of 35 putative GAL4 dimerization domains. GAL4-like transcription factors are extremely frequent in yeast (extrapolations from the available yeast data predict a total of more than 100 [1•]), but have not been found outside the fungi [1•]. They are characterized by the presence of an N-terminal DNA-binding domain. This figure documents the presence of a second conserved motif in many of the GAL4-like proteins; a similarity in this region has already been noted in six closely related members of the GAL4 family [78]. The motif with its two conserved hydrophobic patches may have a role in dimerization of the GAL4-like transcription factors. In the first column are the names of the domains (SWISS-PROT codes); the second column shows positions of the domains in the sequences; and the final column gives database accession numbers (for SWISS-PROT entries, SWISS-PROT numbers are given, otherwise EMBL/Genbank numbers are used). Conserved residues are shown in an outlined larger typeface and conserved hydrophobic positioi.. are shown in bold. Numbers in parentheses indicate omitted residues, and dashes represent gaps (insertions/deletions). The following species abbreviations are given: NEUCR, *Neurospora crassa*; ASPOR, *Aspergillus oryzae*; SCHPO, *Schizosaccharomyces pombe*; EMENI, *Emiricella nidulans*; KLULA, *Kluyveromyces lactis*; ASPNI, *Asparagillus nidulans*; LENED, *Lentinus edodes*.

An improvement of the Gibbs method that allows both identification and statistical evaluation of blocks (motifs) in a large set of protein sequences, and subsequent matrix-type database screening, has been recently published [18].

Despite the high sensitivity of motif searches, the additional use of complementary methods such as profile searches, but also standard database searches, is always recommended, and iterations as well as alternations of different methods are frequently required. In 1995, some progress was reported in the automation of motif identification and database searches [19,20], even though this important direction has not yet led to robust searching machines suitable for general use. The integration of motif analysis in genome scale projects and the generation of genome-specific motif collection is another line of research [21•,22].

Motif discovery in different functional protein classes

Proteins can be classified into several broad functional categories. Often these functions are performed in different compartments of the cell and it appears that motifs are frequently restricted to proteins that share at least one common subfunction (e.g. nucleotide binding). In the following sections we review new motif discoveries classified by the functional categories of the respective proteins. In the course of our studies, we came to realize that a considerable and alarming fraction of the reported 'new' motifs have already been published before. The following collection of motif discoveries published in 1995 can at least be considered 'double-checked'. On the other hand, we certainly cannot guarantee that all new motifs have been included.

Metabolic enzymes are probably the best-studied proteins. They are typically fairly conserved in evolution, and well-suited for motif approaches because of their frequently invariant active site residues. Many motifs from this category of proteins have already been described, and they comprise a considerable fraction of those in PROSITE. The majority of metabolic enzymes bind nucleotides, but their binding sites vary greatly. Thus, it is no surprise that four novel nucleotide-binding motifs or strong deviations from previously known nucleotide-binding sites have been discovered [23,24,25•,26•]; each of the novel motifs is a signature for a whole family of rather divergent yet functionally similar enzymes.

Conserved regions in bacterial toxin ADP-ribosyltransferases were observed in the GPI-linked muscle ADP-ribosyltransferase, a similarity that has been further supported by site-directed mutagenesis [23].

A novel flavin adenine dinucleotide (FAD)-binding motif has been found in a group of bacterial and eukaryotic FAD-dependent oxidases; unexpectedly, the family also included the product of the plant developmental gene *DIMINUTO* [24]. Even though the new motif somewhat resembled the P-loop, it was readily distinguishable from known families of nucleotide-binding proteins by a matrix-based method for motif search.

In order to predict the topology of TagD, a bacterial glycerol-3-phosphate cytidyltransferase, motif searches have identified a diverse group of proteins that share a nucleotide monophosphate (NMP)-binding site but seem to possess at least 10 distinct catalytic activities. An even more remote similarity to class I aminoacyl-tRNA synthetases (which bind adenylate) was then used to build a rather precise 3D model of TagD [25•].

As already mentioned above, a new aspect of motif analysis that is rapidly gaining momentum is the inclusion of 3D information in the strategy of motif delineation. The structural superposition of DNA polymerase β with kanamycin nucleotidyltransferase uncovered not only a common topology but also a similarity in the nucleotide-binding mode [26•]. Motif and profile searches based on the derived sequence alignment revealed an ancient nucleotidyltransferase superfamily that contains different enzymatic activity, including DNA-nucleotidyl exotransferase, Poly(A) polymerase, and glutamine synthase adenylyltransferase [26•].

A surprising evolutionary and functional link between a metabolic enzyme, glycogen phosphorylase, and a DNA-modifying enzyme, T4 β-glucosyltransferase, has been revealed by the superposition of their 3D structures [27]. Although not a single functional residue is identical, there are striking similarities in the spatial arrangement and the chemical nature of the substrates. This similarity has been used to deduce yet another possible relative, namely, T4 α-glucosyltransferase [27].

The translation machinery is generally conserved between eukaryotes and prokaryotes although there are also considerable differences, for example, in the architecture of the ribosome. Yeast genome sequencing revealed the conservation of three translation-associated proteins between eukaryotes and bacteria: namely, peptidyl-tRNA hydrolase, ribosome recycling factor, and a putative translation activator [28].

Sequence analysis of the eukaryotic guanine nucleotide exchange translation initiation factor eIF-2B subunits revealed three unexpected motifs, illustrating the different types of motif discoveries on a single group of proteins [29•]. A new motif is located at the C termini of one of the eIF-2B subunits, two other translation initiation factors (eIF-4g and eIF-5), and an uncharacterized human protein; this motif was implicated in the interaction of each of these proteins with eIF-2. A putative nucleotide-binding domain, which contains a significantly modified P-loop, has been identified in the N-terminal portions of two eIF-2B subunits and a number of nucleotidyltransferases. This domain is likely to be directly involved in the GTP/GDP exchange catalyzed by eIF-2B. Finally, a repetitive motif called the 'isoleucine patch' was detected downstream of the nucleotide-binding domain; this motif is shared by two eIF-2B subunits and a number of nucleotidyltransferases and acetyltransferases. The isoleucine patch may be involved in acyl group binding by acetyltransferases but its role in translation factors remains enigmatic.

Protein folding is tightly associated with translation. In spite of the intense research in this area and the apparently limited number of proteins involved, two new motifs were published in 1995 [30,31].

The first of these motifs defines a new family of peptidylprolyl *cis–trans* isomerases (PPIs) generalized from the experimentally characterized *Escherichia coli* PpiC protein, also called parvulin (see [30] and references therein). The new family brings together several proteins, for which a chaperone-like function had been observed previously but no PPI activity had even been suspected; an example is the NifM protein involved in nitrogen fixation.

The second motif represents a unique identifier for the whole 10 kDa co-chaperonin family [31]. This work shows how sometimes the addition of a single highly divergent new member, in this case the bacteriophage T4 chaperonin, allows the delineation of a motif that could not be detected previously because of the high sequence conservation in the chaperonin family.

Protein splicing is a recently discovered mechanism that excises regions from an already translated polypeptide. A motif that is conserved in the proximal extein–intein junction of such splice sites [32] was later generalized and detected in the hedgehog family of vertebrate and insect morphogens, thus revealing the widespread occurrence and importance of the cleavage mechanism involved in protein splicing [33]. On the basis of this finding, the catalytic cysteine involved in the peptide bond cleavage has been predicted [33]. In a complementary experiment, it has been demonstrated that this cysteine residue is indeed indispensable for hedgehog autoproteolysis [34].

Transcription factors and other proteins involved in gene expression regulation are subject to numerous studies, and secondary functions such as roles in malignancy lead to very different characterization levels of the proteins involved.

A conserved motif that unifies two large, important families of transcription factors, heat-shock factors and ETS proteins, has been described recently [35•]. It is notable that even though the 3D structure has been determined for representative proteins from each of the families, the common motif was found not by their superposition but by a matrix method for motif search. This example shows that even with the increased use of structural information, sophisticated sequence-based methods for motif detection remain competitive.

An example of new, apparently important, functional implications from the analysis of an old motif by more sensitive profile and motif-search methods is the chromo domain discovered as early as 1991 in negative transcription regulators, which is involved in position-effect variegation. Two independent recent studies showed that some of the chromo domain proteins contain a second, divergent copy of the domain dubbed the 'chromo shadow' domain [36]. In addition, the chromo motif was discovered in several new proteins, notably the retinoblastoma-binding protein RBP-1 and the *Drosophila* protein MSL-3 involved in X chromosome dosage compensation [37]. The latter generalization suggests that the chromo domain is a general purpose vehicle for delivering both negative and positive transcription regulators to the sites of their action on chromatin.

Several groups have been actively involved in the systematic analysis of domains in large modular proteins. In particular, this analysis has resulted in the identification of the forkhead associated (FHA) domain in non-DNA binding regions of transcription factors, such as forkhead, and in other putative nuclear transcription-associated proteins [38].

The cell cycle-dependent expression of proteins is a first indication of their involvement in regulation and manifestation of the different stages of the cell cycle. The cloning and sequencing of the cell- and stage-specific murine gene *tbc1* revealed the conservation of a 200 amino acid domain present in other cell cycle proteins such as tre-2, BUB2 and cdc16. This domain has been called TBC and has been proposed to be involved in protein–protein interactions leading to cell-cycle regulation [39].

The cloning and sequencing of Ran/TC4-binding proteins, as well as mutation and deletion analysis, resulted in the characterization of a minimal binding domain that is also present in several other eukaryotic proteins [40]. Ran/TC4 is a nuclear GTPase implicated in the initiation of DNA replication, entry into and exit from mitosis, and nuclear RNA and protein transport. Thus, the new Ran-binding domain may have important regulatory functions in these processes.

A particularly instructive example of an extremely widespread diverse motif is the histone fold, which has been greatly expanded in recent studies [41•,42]. The histone fold motif that was originally derived from the multiple alignment of the four core histone classes was used for extensive database screening by the MoST method [41•]. As a result, the histone fold has been tentatively identified in a variety of transcriptional regulators and other DNA-binding proteins. Additional analysis was needed to filter for false positives. A further generalization of the histone fold has been achieved by the identification of common topologies between histones, LexA, forkhead and the replication terminator proteins [43], although this observation is hard to complement by sequence similarity analysis.

RNA- and DNA-binding domains can be found in a variety of cytoplasmic and nuclear proteins that perform very diverse overall functions. Numerous distinct binding motifs exist but the majority can be grouped together under the term 'zinc finger' (for a recent review see [44]), that is, metal coordination centers that contain at least four cysteine or histidine residues that bind the metal that stabilizes the tertiary structure and mediates binding to DNA or RNA [44]. The different classes of zinc finger may look similar in sequence but can have totally different tertiary structures. On the other hand, divergent variants of a family with a common tertiary structure are sometimes grouped into a separate class. It is difficult to judge reports on a new zinc finger type just by the sequence similarities; at least five distinct variants were reported in 1995 ([45–47,48•]; see also below).

The competition in the field of sequence analysis can be demonstrated by the identification of a new putative zinc finger like DNA binding domain involved in chromatin-mediated transcription control, the PhD finger [49]. As many as three other reports dealt with the delineation of this domain in 1995 [50–52] and, consequently, different names have been introduced; a frequent and difficult problem in the field of motif analysis.

The double stranded RNA binding domain from the *Drosophila* staufen protein has been shown, by structural comparison and subsequent sequence analysis, to be homologous to the N-terminal domain of ribosomal protein S5 [53], thus generalizing this type of RNA-binding domain.

Cytoplasmic regulatory domains have received a lot of attention in the last few years. However, the focus was mainly on signaling cascades, and even here, the SH2, SH3 and PH domain were dominating. Only very recently, many more domains involved in signaling and other regulatory cascades have been identified, and, consequently, a much more complex picture has emerged.

In 1995, considerable progress was made in the identification of proteins and particular domains involved in apoptosis, the programmed cell death. The first of these, the DEATH domain, was already identified in 1992 as a region of similarity between the tumor necrosis factor (TNF) receptor and Fas/Apo1, which also binds TNF-like proteins (for a review see [54•]). Numerous reports on DEATH domains in 1995 revealed a distribution not restricted to apoptosis (see, for example, [54•]). This is a clear example of an important generalization of an existing motif, because nearly 20 distinct cytoplasmic proteins are now known to contain the DEATH domain.

Two unrelated motifs were discovered in 1995 that bind to the DEATH domain in TNF receptor like proteins. The first motifs were found in the mammalian c-IAP1 and 2 and the *Drosophila* DIAP1 and 2, which contain a repeated domain (BIR, baculovirus IAP [inhibitor of apoptosis protein] repeat) with similarity to virus-encoded inhibitors of apoptosis [55,56].

Another group of proteins bind to the DEATH domain of TNF receptor like proteins. One of them, CRAF1 [57] or, alternatively, CAP1 [58], led to the characterization of the minimal binding domain by mutation experiments. This domain (which is unrelated to BIR) is homologous to two other TNF-receptor binding proteins, TRAF1 and TRAF2 [57,58]. The comparison of the proteins belonging to the TNF-receptor associated family resulted in the delineation of three distinct motifs: firstly, a modified RING finger in the N-terminal region; secondly, an original cysteine-rich motif designated CART (C-rich motif associated with RING and TRAF domains) in the middle of the proteins; and thirdly, the C-terminal TRAF domain that is directly associated with the cytoplasmic domain of TNF-receptors [48•].

A new domain that is common to sexual differentiation proteins such as byr2, STE11, STE4 and STE50, and other putative signaling proteins, has been named SAM, for sterile alpha motif [59]. The four predicted helices of the SAM domain may form a bundle analogous to many other recognition proteins.

A common cytoplasmic juxtamembrane domain has been identified in several divergent type II cytokine receptors such as the interferon (IFN)-γ receptor, the interleukin (IL)-10 receptor and tissue factor [60]. This is analogous to a similarly located, functionally important domain shared by many type I cytokine receptors, even though the motifs themselves are unrelated. Thus, the signaling mechanism might be analogous in the two types of cytokine receptors.

The identification of the WW domain present in many cytoplasmic regulatory proteins allowed the specific expression of the domain and its characterization. The subsequent identification and characterization of its ligands revealed a peptide-binding capability for WW. The 3D structure solved for a WW domain bound to the identified proline-rich peptide (H Oschkinat, personal communication) revealed a novel fold and verified that the WW domain has a role analogous to the proline-rich peptide-binding SH3 (for a review see [61]). Coincidentally, a domain that binds phosphotyrosine (PI/PTB) that is analogous to, but structurally different from, SH2, has been found in several signaling proteins using motif searches [62]. Thus, signaling via SH2 and SH3 seems to be only a small component of a large network of interacting proteins and domains.

In dystrophin, a well-characterized large protein, not only the WW domain , but also a cysteine-rich domain called ZZ has been discovered recently; ZZ has been proposed to represent yet another Zn^{2+} coordination center [63].

Another domain that occurs in cytoskeletal proteins has been named CH (calponin homology); it is present in signaling proteins such as Vav which are involved in the activation and inactivation of small G-proteins. Another common feature of CH seems to be its actin-binding capability; the domain is exclusively located at or near the N termini in all proteins for which it has been described so far [64].

Other proteins associated with the cytoskeleton are motor proteins that use ATP to perform directional locomotions along the filament. In members of two such families, namely, myosins and kinesins, regulatory domains have been discovered [65•]. Interestingly, both the regulatory domains found in myosins (DIL) and those found in kinesins (U104) have also been detected in proteins without motor domains, namely, human Af6 and drosophila cno [65•].

Another group of cytoplasmic proteins is involved in transport processes. Surprisingly, the analysis of huntingtin, the product of the gene responsible for Huntington's disease, revealed internal successive repeats (dubbed HEAT) that have been found in diverse cytoplasmic proteins involved in vesicle-associated transport [66]. The group of proteins containing such repeating units, each about 40 residues long, seems to be much larger than anticipated, because there is a considerable similarity between HEAT and ARM repeats (E Hartmann, P Bork, unpublished data) originally found in the armadillo protein [67]. Only the phasing of the two predicted α helices appears to be different, which may be due to the difficulties in determining the domain boundaries in divergent repetitive units.

Extracellular proteins have been known for a long time to consist of modules, and many of them have been extensively classified (see [68] and references therein).

Although the motif discovery seems to slow down in these proteins, quite a few novel domains were reported in 1995.

One of these has been called SEA (first found in a sperm protein, agrin and enterokinase; [69]). It is contained in well-studied proteins such as the matrix protein perlecan, the digestion initiator enterokinase and the synaptic protein agrin. All the proteins containing the SEA modules appear to be heavily O-glycosylated [69]. The discovery was triggered by newly sequenced genes that allowed the delineation of motifs and their identification in well-studied proteins.

Collagens share many modules with other extracellular proteins. The identification of three alternative transcripts for the N-terminal ends of type XVIII collagen and subsequent analysis revealed the presence of a novel cysteine-rich domain that previously has been found only in the extracellular parts of frizzled proteins, G-protein-coupled membrane receptors that are needed for the establishment of cell polarity in the epidermis [70]. This so-called fz domain is characterized by 10 conserved cysteines and might bind to similar ligands in both collagen VIII and frizzled proteins.

Natural killer lysins, effector molecules identified in cytotoxic lymphocytes, were unexpectedly found to structurally and functionally resemble amoebapore, a polypeptide from *Entamoeba histolytica* that forms ion channels in target cell membranes [71]. This similarity allowed the author to propose a common functional mechanism and shows the widespread use of the toxins.

The product of the gene that is mutated in polycystic kidney disease (PKD) contains numerous extracellular domains. PKD1 seems to contain at least 14 copies of a novel domain, although it seems topologically similar to immunoglobulins [72•]. This domain is found not only in other human extracellular proteins but also in several prokaryotic extracellular proteases and archaeal multilayer proteins. So far, among the bacterial proteins, only glycohydrolases have been shown to possess a modular architecture.

The sequencing of a bacterial endo-1,4-β-D-xylanase (XynA) revealed the presence of several non-catalytic domains that were systematically tested for cellulose binding. Two repeated C-terminal domains showed high activity, and subsequent database searches identified this new cellulose-binding domain in various other bacterial extracellular glycohydrolases [73].

Systematic motif discovery as a part of genome analysis

As discussed above, recent literature is replete with modifications and generalizations of already known motifs, but the discovery of really new motifs is relatively infrequent. Evidently, at least three factors contribute to this situation: the objective saturation of the number of known motifs; a lack of systematic effort to discover new motifs; and the inadequacy of the available methods for the discovery of more subtle motifs. In order to evaluate the relative contribution of each of these trends, it is of interest to track down the discovery of new motifs in systematic analyses of large protein ensembles. As a part of one such study, 2328 *E. coli* proteins (about 60% of all gene products) were clustered by sequence similarity to one another, and motifs conserved in each of these clusters of paralogs were systematically explored [20]. Altogether, 166 motifs were delineated. The majority of these motifs are already known, even though in most cases, additional members of the respective protein families were detected; 10 motifs appeared to be new.

Another systematic study focused on the 1703 putative proteins encoded in the complete genome of *Haemophilus influenzae* [74]. In the course of this analysis, 46 motifs were discovered that have not been described previously, yet are conserved, at least at the level of distantly related bacteria. The overall number of conserved motifs present in *H. influenzae* proteins is difficult to estimate. The upper boundary is about 800, as this is the number of distinct conserved regions shared with proteins from phylogenetically distant organisms. This study clearly indicates that most of the motifs are already known but the number of new ones detected by the systematic analysis of all protein sequences encoded in a genome is still considerable.

We present here two out of the many examples of new motifs detected in the course of genome scale sequence analysis. The first of these motifs (Fig. 2a) was originally derived from a Blastp search output [11] for the uncharacterized *H. influenzae* protein YrdC and initially included only other uncharacterized sequences from bacteria and yeast. The subsequent search with the matrix-searching technique (MoST) method showed that this motif is contained also in HypF proteins from various bacteria that are involved in transcription regulation of the hydrogenase operon. The motif shown in Figure 2a does not contain any invariant residues but includes a number of positions occupied by similar (hydrophobic or charged) residues. Thus it is unlikely that this is an active site of an enzyme; rather, this motif may belong to a novel DNA-binding or a protein–protein interaction domain.

The second motif (Fig. 2b) is highly conserved in a number of bacterial and yeast proteins, none of which has been functionally characterized. Among them are four proteins from both *E. coli* and *H. influenzae*, and a protein from *Mycoplasma genitalium* (having one of the smallest genomes of living cells, with only 468 genes). The motif contains an invariant histidine residue preceded by a hydrophobic region (Fig. 2b). It may be speculated that proteins containing this motif possess an enzymatic activity that is essential for any cell but at present we have no clue as to what this activity might be.

The two examples described represent 'motifs in search of function' that are typically discovered in the course of genome sequence analysis. Our study of the *H. influenzae* genome revealed 25 such motifs [74]. Subsets of both protein families discussed above have been independently used by A Bairoch and KE Rudd (personal communication) to derive PROSITE signatures (PS01147 for the first family, and PS01137, PS01090, and PS01091 for the second family), demonstrating once again the redundancy in motif discovery by different groups.

How many motifs are still to be discovered?

The previous paragraph already gave a first clue as to the proportion of detectable sequence motifs in bacteria. Eukaryotes, however, have numerous regulatory pathways that contain proteins with no relatives in prokaryotes. For example, as many as 40% of the mammalian proteins might be (at least partially) extracellular [67]; most of

Figure 2

```
(a)
HYPF_ECOLI    213: GKIVAIKGIGGFHLACDARNSNAVATLRARKHR  P30131
HYPF/SYNSP    213: GNIIAIKGLGGFHLCCDATDFEAVEKLRLRKHR  D64000
HUPY/AZOVI    196: GEIVALRGVGGFHLACDARNAGAVALLRRRKRR  JN0648
HYPF_RHOCA    206: GEILAVKGLGGFHLACDATNADAVDLLRARKRR  Q02987
HYPF_AZOVI    196: GEILALRGVGGFHLACDARNAGAVAELRRRKRR  P40596
HYPF_RHILV    209: GAIVALKGVGGFHLLCDARNDGAIGLLRLRKAG  P28156
YRDC_HAEIN     13: NQVVAYPTEAVFGLGCNPQSESAVKKLLDLKQR  P44807
YWLC_BACSU     31: NEVVAFPTETVYGLGANAKNTDAVKKIYEAKGR  P39153
YRDC_ECOLI      6: ERVIAYPTEAVFGVGCDPDSETAVMRLLELKQR  P45748
YCIO_ECOLI     39: GGVIVYPTDSGYALGCKIEDKNAMERICRIRQL  P45847
YCIO_HAEIN     27: GGVIVYPTDSGYALGCMMGDKHAMDRIVAIRKL  P45103
YRFE_MYCLE     29: GRLVVMPTDTVYGIGADAFDRAAVAALLSAKGR  P45831
SUA5_YEAST     63: DETVAFPTETVYGLGGSALNDNSVLSIYRAKNR  P32579

(b)                                 *
YG64_HAEIN    113: LERFILIAKKWDLPLNLHIVHNDVEIALELL  P45305
YJJV_HAEIN    121: FESQLYLAKQFNLPVNIHSRKTHDQIFTFLK  P44500
YIGW_ECOLI    114: FVAQLRIAADLNMPVFMHCRDAHERFMTLLE  P27859
YJJV_ECOLI    116: LDEQLKLAKRYDLPVILHSRRTHDKLAMHLK  P39408
YABD_BACSU    111: FRNQIALAKEVNLPIIIHNRDATEDVVTILK  P37545
Y009_MYCGE    115: FEMQFEIAETNKLVHMLHIRDAHEKIYEILT  P47255
YCFH_HAEIN    114: FGSQIDIANQLDKPVIIHTRSAGDDTIAMLR  P44718
YTP3_YEAST    163: FRRFCRLARHTSKPISIHDVKCHGKLNDICN  P38430
YBF5_YEAST    208: LKISCLNDKLSSYPLFLHMRSACDDFVQILE  P34220
SCN1_SCHPO    220: FEAQVRLAAEFQRAVSVHCVQTYALLYSSLA  P41890
```

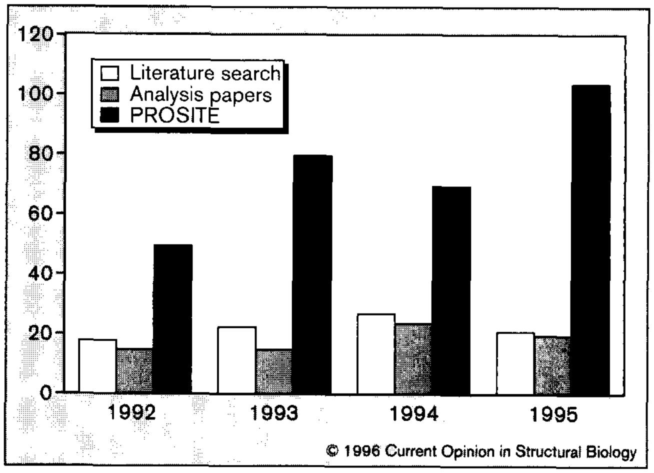
Examples of motifs detected in the course of systematic genome sequence analysis. In the first column are the names of the domains (SWISS-PROT codes); the second column shows positions of the domains in the sequences; and the final column gives database accession numbers (for SWISS-PROT entries, SWISS-PROT numbers are given, otherwise EMBL/Genbank numbers are used). Conserved residues are shown in an outlined larger typeface and conserved hydrophobic positions are shown in bold. The species abbreviations are the same as those used in Figure 1, along with the following: ECOLI, *Escherichia coli*; SYNSP, *Synechococcus sp*; AZOVI, *Azotobacter vinelandii*; RHOCA, *Rhodobacter capsulata*; RHILV, *Rhizobium leguminosarum*; HAEIN, *Haemophilus influenzae*; BACSU, *Bacillus subtilis*; MYCLE, *Mycobacterium leprae*; MYCGE, *Mycoplasma genitalium*. **(a)** A putative binding motif in HypF transcription regulators and uncharacterized bacterial and yeast proteins. **(b)** A highly conserved putative catalytic motif in uncharacterized bacterial and eukaryotic proteins. The alignment is the MoST program output, to which the YBF5 sequence was added on the basis of additional Blastp searches. The putative catalytic histidine is indicated by an asterisk.

them contain conserved disulfide bridges detectable as characteristic cysteine motifs.

To get a rough estimate of the speed and possible saturation of motif identification, Figure 3 shows the result of a literature search from 1992–1995, a summation of the papers published in journal sections devoted to sequence-similarity analysis and motif discovery, and the growth of the PROSITE sequence pattern library [3•]. Although the numbers are only approximate, they suggest that there is still linear growth in the discovery of new motifs. As we have excluded from our statistics numerous motifs discovered in the course of genome analysis, this tendency seems somewhat at variance to proposals of saturation of the number of motifs to be discovered [75]. Several facts might account for the discrepancy: the speed of sequencing is increasing further, which may prevent gradual saturation but will subsequently result in a rather abrupt end to motif discoveries; many motifs are specific to a particular phylogenetic division (e.g. Metazoa), therefore, as soon as the representation of this division in sequence databases improves significantly there will be a boost to motif discovery; the methods for motif discovery have improved, such that motifs have been found that could not be detected previously. This includes the consideration of similarities of 3D structures as an initial step in motif definition (see, for example, [76]).

Thus, the discussion about the number of motifs is comparable to the one on the number of protein folds [77]: there is a limited number of widespread motifs/folds, of which we already know the majority, but there is also a 'tail' of numerous motifs/folds that will continue to be identified, even after complete genome sequences for several organisms become available.

Figure 3



© 1996 Current Opinion in Structural Biology

Statistics on motif identification. In order to judge the number of motifs (vertical axis) identified within the last few years, we performed some independent estimates. A keyword literature search (white bars), including subsequent manual filtering, was complemented by a detailed inspection of journals in which motif discoveries without experimental work are frequently described (e.g. *Cell, Trends in Biochemical Sciences, Protein Science, Nucleic Acids Research*; shaded bars). There was only a little overlap between the two. The detailed compilation for this article revealed a total of 35 motif discoveries for 1995 (the criteria for inclusion are given in the introduction). Thus, the numbers given can only be considered as a very rough estimate. The black bars indicate the motifs that are newly entered into PROSITE. Note that in PROSITE, various motif types are included (see introduction) that we have not considered here. Nevertheless, all three independent measurements indicate more or less linear growth, rather than a saturation, in the discovery of new motifs.

Conclusions

Protein motif identification has become an essential part of sequence analysis in general and genome research in particular. A somewhat surprising tendency is that new motifs are still being discovered at about the same pace as a few years ago, but an increasing number of reports now deal with the detection of an already known, although frequently modified, motif in another group of proteins. Another important trend is the unification of two or more motifs leading to a single, more general signature that allows protein families to be grouped into superfamilies. The comparison of different 3D structures and sequence–structure comparisons are becoming increasingly important as the number of available structures grows rapidly, and the methods for their analysis improve. Nevertheless, methods for sequence-motif detection remain complementary to structure-based methods.

The competition in the field and the increasing complexity of exhaustive literature searches lead to numerous cases of simultaneous motif discoveries by independent groups and to problems in assigning unique names to particular motifs/protein families. In addition, a not insignificant fraction of the motif discoveries turn out to be statistically unsound. The further digestion of all the information, especially in the light of systematic genome analysis, requires a concerted effort in the identification of all existing motifs. Expert annotation as found in the PROSITE database should be complemented with automatic delineation and classification of motifs.

Acknowledgements

We apologize to all those researchers who have published new motifs that meet our definition but have not been included here. Such findings are often hidden in the text and not retrievable by automated literature searches. Special thanks go to our co-authors on various reports of new motifs and methods for motif identification. We are indebted to Rein Aasland for information on the PhD fingers, Amos Bairoch for the PROSITE statistics, Erik Sonnhammer and Hartmut Oschkinat for the communication of unpublished results, as well as Rebecca Wade, Nigel Brown, Jörg Schultz and Hedvig Hegyi for critical reading of the manuscript.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

• of special interest
•• of outstanding interest

1. • Bork P, Gibson TJ: **Applying motif and profile searches.** *Methods Enzymol* 1996, **266**:162–184.

This paper constitutes a critical overview of most of the existing methods for motif and profile analysis and a guide as to their applications.

2. Bork P: **Mobile modules and motifs.** *Curr Opin Struct Biol* 1992, **2**:413–421.

3. • Bairoch A, Bucher P, Hofmann K: **PROSITE: new developments.** *Nucleic Acids Res* 1996, **24**:189–196.

PROSITE is a valuable source for well-annotated protein sequence motifs. An additional (profile) description for motifs has been introduced that complements the existing patterns.

4. Pietrokovski S, Henikoff JG, Henikoff S: **The BLOCKS database – a system for protein classification.** *Nucleic Acids Res* 1996, **24**:197–200.

5. Attwood TK, Beck ME, Bleasby AJ, Degtyarenko K, Smityh DJP: **Progress with the PRINTS protein fingerprint database.** *Nucleic Acids Res* 1996, **24**:182–183.

6. Murvai J, Gabrielian A, Fabian P, Hatsagi Z, Degtyarenko K, Hegyi H, Pongor S: **The SBASE protein domain library, release 4.0: a collection of annotated protein sequence sgments.** *Nucleic Acids Res* 1996, **24**:210–213.

7. Fitch W: **Distinguishing homologous from analogous proteins.** *Syst Zool* 1970, **19**:99–106.

8. Doolittle RF: **Convergent evolution: the need to be explicit.** *Trends Biochem Sci* 1994, **19**:15–18.

9. Holm L, Sander C: **Searching protein structure databases has come of age.** *Proteins* 1994, **19**:165–173.

10. • Bork P, Ouzounis C, McEntyre J: **Ready for a motif submission? A proposed checklist.** *Trends Biochem Sci* 1995, **20**:104.

This paper provides a draft of a minimal set of rules making a new motif admissible.

11. Altschul SF, Boguski MS, Gish W, Wootton JC: **Issues in searching molecular sequence databases.** *Nat Genet* 1994, **6**:119–129.

12. Claverie JM: **Some useful statistical properties of position-dependent weight matrices.** *Comput Chem* 1994, **18**:287–294.

13. Neuwald AF, Green P: **Detecting patterns in protein sequences.** *J Mol Biol* 1994, **239**:698–712.

14. Wang JT, Marr TG, Shasha D, Shapiro BA, Chirn GW: **Discovering active motifs in sets of related proteins sequences and using them for classification.** *Nucleic Acids Res* 1994, **22**:2769–2775.

15. Nakata K: **Prediction of zinc finger DNA binding protein.** *CABIOS* 1995, **11**:125–131.

16. Sagot MF, Viari A, Pothier J, Soldano H: **Finding flexible patterns in a text: an application to three-dimensional molecular matching.** *CABIOS* 1995, **11**:59–70.

17. Tatusov RL, Altschul SF, Koonin EV: **Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks.** *Proc Natl Acad Sci USA* 1994, **91**:12091–12095.

18. Neuwald AF, Liu JS, Lawrence CE: **Gibbs motif sampling: detection of bacterial outer membrane protein repeats.** *Protein Sci* 1995, **4**:1618–1632.

19. Bailey TL, Elkan C: **The value of prior knowledge in discovering motifs with MEME.** *Intelligent Systems in Molecular Biology* 1995, **3**:21–29.

20. Wu TD, Brutlag DL: **Identification of protein motifs using conserved amino acid properties and partitioning techniques.** *Intelligent Systems in Molecular Biology* 1995, **3**:402–410.

21. • Koonin EV, Tatusov RL, Rudd KE: **Sequence similarity analysis of *Escherichia coli* proteins: functional and evolutionary implications.** *Proc Natl Acad Sci USA* 1995, **92**:11921–11925.

A prototype of a genome-scale sequence analysis study that includes motif analysis as an integral component, with a specialized motif library being one of the results.

22. Koonin EV, Tatusov RL, Rudd KE: **Protein sequence comparison at genome scale.** *Methods Enzymol* 1996, in press.

23. Takada T, Iida K, Moss J: **Conservation of a common motif in enzymes catalysing ADP-ribose transfer.** *J Biol Chem* 1995, **270**:541–544.

24. Mushegian AR, Koonin EV: **A putative FAD-binding domain in a distinct group of oxidases including a protein involved in plant development.** *Protein Sci* 1995, **4**:1243–1244.

25. • Bork P, Holm L, Koonin EV, Sander C: **The cytidylyltransferase superfamily: identification of the nucleotide-binding site and fold prediction.** *Proteins* 1995, **22**:259–266.

An example of the extension of a motif initially derived by sequence analysis into the sequence–structure comparison area, resulting in a non-trivial connection between various nucleotidyltransferases and aminoacyl t-tRNA synthetases.

26. • Holm L, Sander C: **DNA polymerase b belongs to an ancient nucleotidyltransferase superfamily.** *Trends Biochem Sci* 1995, **20**:345–347.

An example of a successful transfer of a discovered structural similarity to a sequence motif, resulting in the detection of a connection between DNA polymerases and nucleotidyl transferases that may have far reaching biological implications.

27. Holm L, Sander C: **Evolutionary link between glycogen phosphorylase and a DNA modifying enzyme.** *EMBO J* 1995, **14**:1287–1293.

28. Ouzounis CA, Bork P, Casari G, Sander C: **New protein functions in yeast chromosome VIII.** *Protein Sci* 1996, **4**:2426–2428.

29. • Koonin EV: **Multidomain organization of eukaryotic guanine nucleotide exchange translation initiation factor eIF-2B subunits revealed by analysis of conserved sequence motifs.** *Protein Sci* 1995, **4**:1608–1617.

The paper describes the detection of a new motif, a dramatic modification of a known motif, and a known motif in an unexpected place in a single protein family.

30. Rudd KE, Rouviere PE, Lazar S, Sofia H, Plunkett G, Koonin EV: **A new family of peptidyl-prolyl isomerases.** *Trends Biochem Sci* 1995, **20**:12–14.

31. Koonin EV, Van der Vies SM: **Conserved sequence motifs in bacterial and bacteriophage chaperonins.** *Trends Biochem Sci* 1995, **20**:14–15.

32. Pietrokovski S: **Conserved sequence features of inteins (protein introns) and their use in identifying new inteins and related proteins.** *Protein Sci* 1994, **3**:2340–2350.

33. Koonin EV: **A protein splice-junction motif in hedgehog family proteins.** *Trends Biochem Sci* 1995, **20**:141–142.

34. Porter JP, Von Kessler DP, Ekker SC, Young KE, Lee JJ, Moses K, Beachy PA: **The product of hedgehog autoproteolytic cleavage active in local and long-range signalling.** *Nature* 1995, **374**:363–366.

35. • Landsmann D, Wolffe AP: **Common sequence and structural features in the heat-shock factor and Ets families of DNA-binding domain.** *Trends Biochem Sci* 1995, **20**:225–226.

This paper demonstrates a biologically important relationship between two families of transcription factors that was not suspected before in spite of the availability of 3D structures for both of them.

36. Aasland R, Stewart AF: **The chromo shadow domain, a second chromo domain in heterochromatin-binding protein 1, HP1.** *Nucleic Acids Res* 1995, **23**:3163–3173.

37. Koonin EV, Zhou S, Lucchesi JC: **The chromo superfamily: new members, duplication of the chromo domain and possible role in delivering transcription regulators to chromatin.** *Nucleic Acids Res* 1995, **23**:4229–4233.

38. Hoffmann K, Bucher P: **The FHA domain: a putative nuclear signalling domain found in protein kinases and transcription factors.** *Trends Biochem Sci* 1995, **20**:347–349.

39. Richardson PM, Zon LI: **Molecular cloning of a cDNA with a novel domain present in the *tre-2* oncogene and the yeast cell cycle regulator *BUB2* and *cdc16*.** *Oncogene* 1995, **11**:1139–1148.

40. Beddow AL, Richards SA, Orem NR, Macara IG: **The Ran/TC4 GTPase-binding domain: identification by expression cloning and characterization of a conserved sequence motif.** *Proc Natl Acad Sci USA* 1995, **92**:3328–3332.

41. • Baxevanis AD, Arents G, Moundrianakis EN, Landsman D: **A variety of DNA-binding and multimeric proteins contain the histone fold motif.** *Nucleic Acids Res* 1995, **23**:2685–2691.

This paper constitutes an example of a very thorough study of a widespread motif based on the systematic application of an iterative MoST. The study demonstrates both the power of this method and the need for cautious interpretation of the results using additional structural and biological criteria.

42. Arents G, Moundrianakis EN: **The histone fold: a ubiquitous architectural motif utilized in DNA compaction and protein dimerization.** *Proc Natl Acad Sci USA* 1995, **92**:11170–11174.

43. Swindells M: **Identification of a common fold in the replication terminator proteins suggests a possible mode for DNA binding.** *Trends Biochem Sci* 1995, **20**:300–302.

44. Klug A, Schwabe JW: **Protein motifs 5. Zinc fingers.** *FASEB J* 1995, **9**:597–604.

45. Stolow DT, Haynes SR: **Cabeza, a *Drosophila* gene encoding a novel RNA-binding protein, shares homology with EWS and TLS, two genes involved in human sarcoma formation.** *Nucleic Acids Res* 1995, **28**:835–843.

46. Weigel D: **The APETALA2 domain is related to a novel type of DNA-binding domain.** *Plant Cell* 1995, **7**:388–389.

47. Yanagisawa S: **A novel DNA-binding domain that may form a single zinc finger motif.** *Nucleic Acids Res* 1995, **23**:3403–3410.

48. • Regnier CH, Tomasetto C, Moog-Lutz C, Chenard MP, Wendling C, Basset P, Rio MC: **Presence of a new conserved domain in CART1, a novel member of the tumor necrosis factor receptor-associated protein family, which is expressed in breast carcinoma.** *J Biol Chem* 1995, **270**:25715–25721.

This paper describes the identification of a modified RING finger and a novel zinc finger-like domain as well as a domain shared with TRAF proteins.

49. Aasland R, Gibson TJ, Stewart AF: **The PHD finger: implications for chromatin-mediated transcriptional regulation.** *Trends Biochem Sci* 1995, **20**:56–58.

50. Koken MH, Saib A, De The H: **A C4HC3 zinc finger motif.** *CR Acad Sci III* 1995, **318**:733–739.

51. Saha V, Chaplin T, Gregorini A, Ayton P, Young BD: **The leukemia-associated-protein (LAP) domain, a cysteine-rich motif, is present in a wide range of proteins, including MLL, AF10, and MLLT6 proteins.** *Proc Natl Acad Sci USA* 1995, **92**:9737–9741.

52. Stassen MJ, Bailey D, Nelson S, Chinwalla, Harte PJ: **The *Drosophila* thrithorax proteins contain a novel variant of the nuclear receptor type DNA binding domain and an ancient conserved motif found in other chromosomal proteins.** *Mech Dev* 1995, **52**:209–223.

53. Bycroft M, Grunert S, Murzin AG, Proctor M, St Johnston D: **NMR solution structure of a dsRNA binding domain from *Drosophila* staufen protein reveals homology to the N-terminal domain of ribosomal protein S5.** *EMBO J* 1995, **14**:3563–3571.

54. • Feinstein E, Wallach D, Boldin M, Vafolomeev E, Kimchi A:**The death domain: a module shared by proteins with diverse functions.** *Trends Biochem Sci* 1995, **20**:342–345.

A concise overview of the remarkable spread of the DEATH motif and its history.

55. Hay AB, Wassarman DA, Rubin GM: ***Drosophila* homologs of Baculovirus inhibitor of apoptosis proteins function to block cell death.** *Cell* 1995, **83**:1253–1262.

56. Rothe M, Pan M-G, Henzel WJ, Ayres TM, Goeddel DV: **The TNFR2–TRAF signaling complex contains two novel proteins related to baculoviral inhibitor of apoptosis proteins.** *Cell* 1995, **83**:1243–1252.

57. Cheng G, Cleary AM, Ye Z, Hong DI, Lederman S, Baltimore D: **Involvement of CRAF1, a relative of TRAF, in CD40 signalling.** *Science* 1995, **267**:1494–1498.

58. Sato T, Itie S, Reed JC: **A novel member of the TRAF family of putative signal transducing proteins binds to the cytosolic domain of CD40.** *FEBS Lett* 1995, **358**:113–118.

59. Ponting CP: **SAM: a novel motif in yeast sterile and *Drosophila* polyhomeotic proteins.** *Protein Sci* 1995, **4**:1928–1930.

60. Mullersman JE, Pfeffer LM: **A novel cytoplasmic homology domain in interferon receptor.** *Trends Biochem Sci* 1995, **20**:55–56.

61. Sudol M, Chen HI, Bougeret C, Einbond A, Bork P: **Characterization of a novel protein-binding module – the WW domain.** *FEBS Lett* 1995, **369**:67–71.

62. Bork P, Margolis B: **A phosphotyrosine-interaction domain.** *Cell* 1995, **80**:693–694.

63. Ponting CP, Blake DJ, Davies KE, Kendrick-Jones J, Winder SJ: **ZZ and TAZ: new putative zinc fingers in dystrophin and other proteins.** *Trends Biochem Sci* 1996, **21**:11–13.

64. Casstresana J, Saraste M: **Does Vav bind to F-actin through a CH domain?** *FEBS Lett* 1995, **374**:149–151.

65. • Ponting CP: **AF-6/cno: neither a kinesin nor a myosin, but a bit of both.** *Trends Biochem Sci* 1995, **20**:265–267.

The author describes the detection of regulatory motifs in motor proteins in an unexpected context.

66. Andrade MA, Bork P: **HEAT repeats in Huntington's disease protein.** *Nat Genet* 1995, **11**:114–115.

67. Peifer M, Berg S, Reynolds AB: **A repeating amino acid motif shared by proteins with diverse cellular roles.** *Cell* 1994, **76**:789–791.

68. Bork P, Bairoch A: **Extracellular protein modules: a proposed nomenclature.** *Trends Biochem Sci* 1995, **20**:poster CO3.

69. Bork P, Patthy L: **The SEA module: a new extracellular domain associated with O-glycosylation.** *Protein Sci* 1995, **4**:1421–1425.

70. Rehn M, Pihilajaniemi T: **Identification of three N-terminal ends of type XVIII collagen chains and tissue-specific differences in the expression of the corresponding transcripts.** *J Biol Chem* 1995, **270**:4705–4711.

71. Leippe M: **Ancient weapons: NK-lysine is a mammalian homolog to pore-forming peptides of a protozoan parasite.** *Cell* 1995, **83**:17–18.

72. • International PKD1 consortium: **Polycystic kidney disease: complete structure of the gene and its protein.** *Cell* 1995, **81**:289–298.

The discovery of an ancient domain in the PKD protein emphasizes the importance of motif analysis for large multidomain proteins; many proteins implicated in human diseases belong to this category.

73. Winterhalter C, Heinrich P, Candussio A, Wich G, Liebl W: **Identification of a novel cellulose-binding domain within the multidomain 120kD xylanase XynA of the hyperthermophile bacterium *Thermotoga maritima*.** *Mol Microbiol* 1995, **15**:431–444.

74. Tatusov RL, Mushegian AR, Bork P, Brown NP, Hayes WS, Borodovsky M, Rudd KE, Koonin EV: **Metabolism and evolution of *Haemophilus influenzae* deduced from a whole genome comparison with *Escherichia coli*.** *Curr Biol* 1996, **6**:279–291.

75. Green P, Lipman D, Hillier L, Waterston R, States D, Claverie JM: **Ancient conserved regions in new gene sequences and the protein databases.** *Science* 1993, **259**:1711–1716.

76. Bork P, Gellerich J, Groth H, Hooft R, Martin F: **Divergent evolution of a beta/alpha-barrel subclass: detection of numerous phosphate-binding sites by motif search.** *Protein Sci* 1995, **4**:268–274.

77. Orengo CA, Jones DT, Thornton JM: **Protein superfamilies and domain superfolds.** *Nature* 1994, **372**:631–634.

78. Chasman DI, Kornberg RD: **GAL4 protein: purification, association with GAL80 protein and conserved domain structure.** *Mol Cell Biol* 1990, **10**:2916–2923.