Recognition of functional regions in primary structures using a set of property patterns

Peer Bork

Academy of Sciences of German Democratic Republic, Central Institute of Molecular Biology, Department of Biomathematics, Robert-Rössle-Str. 10, Berlin 1115, Germany

Received 25 August 1989

32 consensus patterns for a set of functional regions and structural motifs in protein sequences were constructed. The pattern definition is heuristic and based on 11 selected steric and physicochemical properties. By comparison with these patterns, it was possible to identify, without false detection, 1532 sites in 8702 protein sequences of SWISSPROT. Screening against such a pattern library offers a considerable chance to identify functional regions or structural motifs in proteins from which only the sequence is known.

Pattern search; Property pattern; Primary structure; Recognition

1. INTRODUCTION

With the progress in sequencing techniques, the body of known primary structures has grown much faster than the number of well-characterized proteins. This tendency is likely to continue in the future, and it presents a challenge to develop theoretical methods for prediction of functional domains and structural features on a heuristic basis.

A popular method to tackle the problem is homology search, but this method fails frequently, and then a further characterization of structure and function is difficult. Since evolutionary selection appears to retain only a limited set of structural principles and functional mechanisms, there is a good chance to predict such features on the basis of known consensus patterns. At present, several groups collect such patterns [1-5]. The heuristic principle as applied is usually derived from the presence of certain amino acids at a given position ('letter coincidence' or 'similarity' of letters), but often the variability in the sequence segments is too high for such a straightforward analysis. Therefore, a number of different approaches has recently been developed [6-11]. Our goal was to describe local regions in terms of biochemical properties. A simple pattern description on the basis of 11 steric and physicochemical properties was surprisingly found to be suitable [12,13]. Currently, we are developing a library of such property patterns of various functional regions and structural motifs. Here we present a set of patterns which can be used for rapid recognition of functional and structural features within protein sequences.

Correspondence address: P. Bork, Academy of Sciences of German Democratic Republic, Central Institute of Molecular Biology, Department of Biomathematics, Robert-Rössle-Str. 10, Berlin 1115, Germany

2. MATERIALS AND METHODS

Following the concept of Taylor [7], for better explanation of evolutionary constraints in a given position of a primary structure, a vector of 11 steric and physicochemical properties (hydrophobic, positively charged, negatively charged, polar, generally charged, small, tiny, aliphatic, aromatic, proline, glycine), considered to be present or non-present, was assigned to each amino acid. Proline as well as glycine has been treated as a self-contained 'property'. Important features such as bulkiness or polarity were further divided into subgroups (e.g. polar, charged, positively charged, negatively charged), at the accepted cost of some redundancy in the property vector.

Our program PAT [12] is able to analyze a given alignment position by position, and it records those properties that are common in all aminó acids and those which never occur in any amino acid found at this position (see fig.1a). Properties which are only sometimes present were recorded, but not used further. In every position of the construct obtained (fig.1a), a characteristic combination of features standing for a set of amino acids (fig.1b) resulted. Amino acids which do not occur in the master set but which show the same combination of properties may also match the pattern. Deletions and insertions are considered and it is possible to combine distant sequence segments of a protein. The number of allowed mismatches can be specified so that the stringency may be 'loosened' (for details see [12,13]).

With the pattern so constructed, sequence databases may be screened (for a review of pattern search methods see [14]). Fig.1c shows an example of recognition of distantly related protein families. The property pattern was constructed from a master set of eucaryotic aspartic proteases (such as pepsin and renin). It turned out that in addition to the expected aspartic proteases, also HIV-proteases fit the pattern as well. A functional and structural relationship between these two protein families is therefore suggested by this finding. With a conventional homology search, this relationship would not have been found. This prediction was made earlier (see e.g. [15]) and it was recently confirmed by X-ray crystallography [16]. For a further fami-

Published by Elsevier Science Publishers B.V. (Biomedical Division)

ly of viral proteins, a similar kinship can be proposed on the basis of the findings (fig.1c).

From several alignments taken from the literature, we extracted property patterns and implemented them in our pattern library. A pattern qualified for inclusion into the library if at least 10 sequence segments in SWISSPROT matched the property pattern correctly. This simple criterion should help to prevent too 'individual' patterns. A stringency criterion was defined such that no obviously incorrect example occurs in SWISSPROT. This strict stringency criterion may be loosened by accpeting more mismatches. Usually, more correct regions were 'redetected' at the risk of false positives (that have nothing to do with the regions characterized by the pattern). An empirical coefficient ('specificity coefficient') for the predictive quality (at a given mismatch number) is the ratio of correct to the total number of detections of that pattern in SWISSPROT. This specificity coefficient will be given for any identified motif in a protein sequence (see fig.3).

3. RESULTS AND DISCUSSION

The 32 property patterns that we have collected may be divided into three groups: (1) ligand binding sites; (2) binding domains of extracellular proteins; and (3) functional regions in members of protein superfamilies. To reduce noise effects, we included only characteristic parts of a domain into a pattern. All 32 patterns described below are listed in fig.2.

Ligand binding sites comprise usually one or more local regions forming a structural motif and containing residues essential for the binding function. Our set of such sites comprises the calcium binding EF-hand motif, two well known DNA binding sites (two zinc finger motif variants and the helix-turn-helix motif), a probable RNA binding site (two conserved motifs in several RNA binding proteins), and the ligand binding sites of proteins that are probably structurally related to β-lactoglobins. The nucleotide binding sites which also belong to this group have been investigated more extensively in [13]. In the present set, binding sites for NAD$^+$, FAD and ATP have been included, as well as a binding site of cyclic nucleotide monophosphate binding, and two GTP binding sites (one for tubulins and the other one for elongations factors, G- and ras-related proteins, etc).

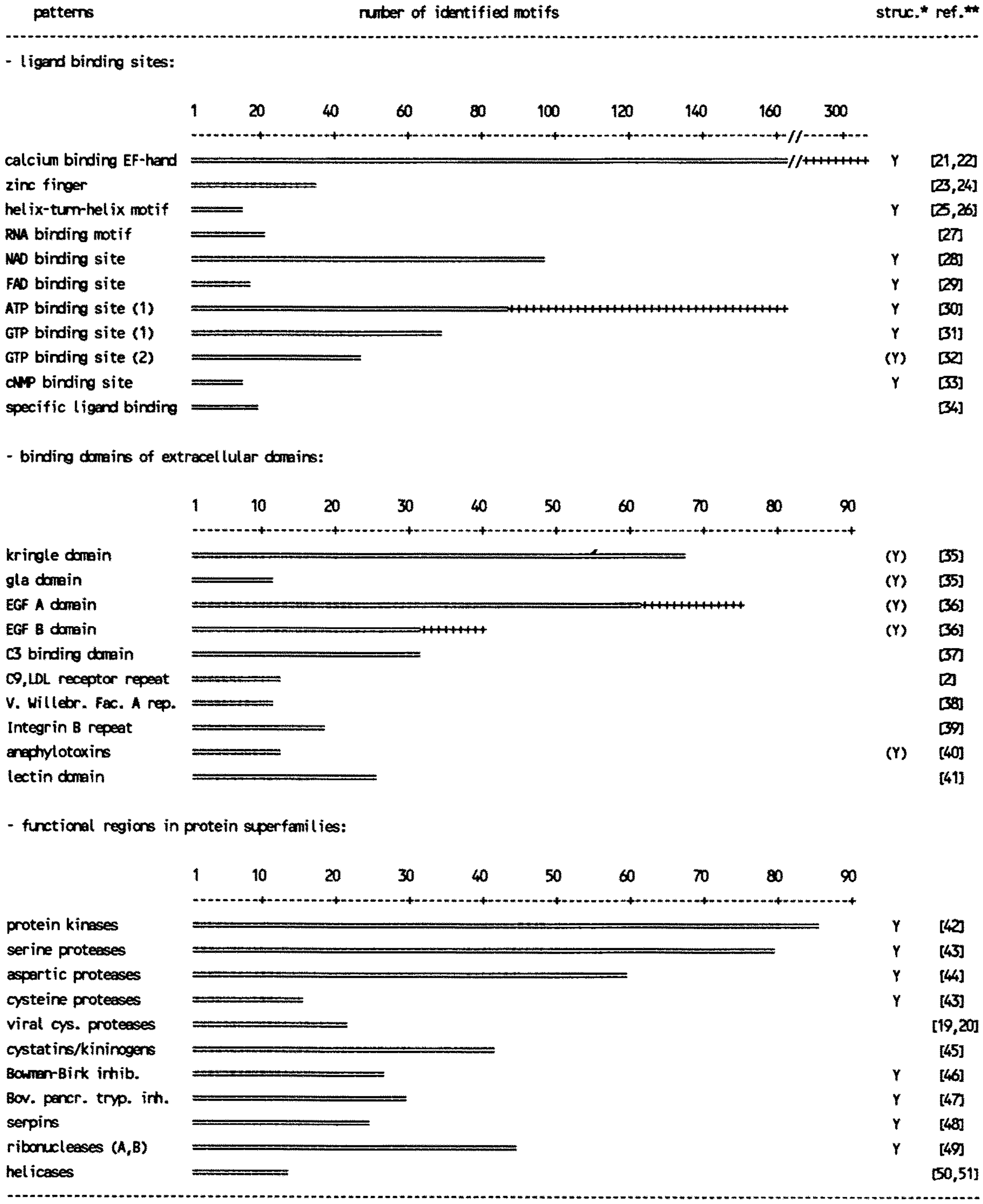
The second group comprises domains obviously assembled by 'exon shuffling' [18]. It turns out that extracellular proteins from the most diverse pathways and with the most disparate functions have surprisingly many domains in common. A complete set of such domains would be a powerful tool in the study of extracellular proteins. At present a large number of such domains are known in more or less detail. Included in our

```
a) remark                      !!!
hydrophobic       .1......1.111011.....1....1.
positive charg.   00.000000.000000000000000.0.
negative charg.   .000....0.0001000..000000.0.
polar             ..........000110.....0......
charged           .0.0....0.0001000..000000.0.
small             ..........1..11111....1.....
tiny              ........0..00001..0.........
aliphatic         .....000...1.00000....00...0
aromatic          00.00000.000.00000..00000.00
proline           0000..0.0000000000000.00.0.
glycine           ...00...00.0000100000.0.000.

b)                AAAAAAAACAAICDTGAACAAAAAAAAA
                  CCCCCCCCFCCLF   CCDCCCCCCCCC
                  DGFIDDDDIDGVI   NDEFIGNGIDID
                  EIGLEEEELEV L   SNFILIPNLELE
                  GLHMIGGGMH  M   TSILMLSSMFMG
                  IMINLMMMTI  V    TLMNMTTNHTK
                  LTKQMNNNVK        MNQV  QIVM
                  MVLSNPQPWL        NQS   SK N
                  N MTPQSQYM        QST   TL P
                  Q NVQSTS N        TTV   VM Q
                  S Q ST T Q        VV     N R
                  T R T    R        WW     P S
                  V S V    S        YY     Q T
                    T      T               R
                    V      V               S
                    W                      T
                    Y                      V
                                           W
                                           Y

c)
CARP$RHIM     94 1 SIGTPGQDFLLLFDTGSSDTWVPhKGCT
             293 5 AAVrfSRPqAftIDTGTNFFIMPSSAAS
CATD$HUMAN    84 0 GIGTPPQCFTVVFDTGSSNLWVPSIHCK
             282 1 GLTLCKEGCEAIVDTGTSLMVGPvDEVR
CATD$PIG      20 0 GIGTPPQCFTVVFDTGSSNLWVPSIHCK
             211 2 SLTLCKGGCEAIVDTGTSLIVGqpEEVR
CHYM$BOVIN    79 0 YLGTPPQEFTVLFDTGSSDFWVPSIYCK
             261 2 VVVACEGGCQAILDTGTSkLVGPSSDIl
PENP$PENJA    20 2 PVTIGGTTLNLnFDTGSADLWVfSTELP
             200 2 AGSQSGDGFSGIaDTGTTLLLLbDSVVS
PEP4$YEAST    96 0 TLGTPPQNFKVILDTGSSNLWVPSNECG
             281 2 DeYAElESHGAAIDTGTSLITLPSGLAE
PEPA$CHICK    64 0 SIGTPQQDFSVIFDTGSSNLWVPSIYCK
             247 3 KyVACffTCQAIVDTGTSLLVMPQGAYN
PEPA$HUMAN    81 0 GIGTPAQDFTVVFDTGSSNLWVPSVYCS
             264 0 EAIACAEGCQAIVDTGTSLLTGPTSPIA
PEPA$MACFU    66 0 GIGTPAQDFTVIFDTGSSNLWVPSVYCS
             249 0 EAIACAEGCQAIVDTGTSLLTGPTSPIA
PEPC$MACFU    62 0 SIGTPPQNFLVLFDTGSSNLWVPSVYCQ
             247 2 AsGwCSEGCQAIVDTGTSLLTVPQQYMS
PEPC$RAT      81 0 SIGTPPQNFLVLFDTGSSNLWVSSVYCQ
             267 0 SGWCSSQGCQGIVDTGTSLLVMPAQYLS
PEPR$RHICH    49 1 TIGTPGKKFNLdFDTGSSDLWIASTLCT
             232 0 GTSTVASSFDGILDTGTTLLILPNNVAA
POL$EQIAV     92 0 IVLINDTPLNVLLDTGADTSVLTTAHYN
POL$HIV10     80 3 TIKIGGQlkEALLDTGADDTVLeEMSLP
POL$HIV1A     68 3 TIRIGGQlkEALLDTGADDTVLeEMNLP
POL$HIV1E     67 3 AIKIGGQlkEALLDTGADDTVLeEMNLP
POL$HIV1M     67 3 TVRVGGQlkEALLDTGADDTVLeEINLP
POL$HIV1P     80 3 TIKIGGQlkEALLDTGADDTVLeEMSLP
POL$HIV1R     67 3 TVKIGGQlkEALLDTGADDTVLeEMNLP
POL$HIV1X     68 3 TIKIGGQlkEALLDTGADDTVLeEMSLP
POL$HIV2R     97 0 TAYIEGQPVEVLLDTGADDSIVAGIELG
POL$MLVAV     14 4 TLTVGGQPVTfLVDTGAqhSVLTQNPgP
POL$MLVMO     14 4 TLKVGGQPVTfLVDTGAqhSVLTQNPgP
POL$SIVAT    121 2 TVYIEGvPIKALLDTGADDTIIkENDLQ
POL$SIVM1    117 0 TAHIEGQPVEVLLDTGADDSIVTGIELG
POL$SIVMK    117 0 TAHIEGQPVEVLLDTGADDSIVTGIELG
POL$VILV      50 4 EIKVGTRwkKLLVDTGADkTIVTShDMS
POLH$DROME    17 1 TIKyKENNLKCLIDTGSTVNMTSKNIFD
RENI$HUMAN    91 0 GIGTPPQTFKVVFDTGSSNVWVPSSKCS
             279 0 STLLCEDGCLALVDTGASYISGSTSSIE
RENI$MOUSE    89 0 GIGTPPQTFKVIFDTGSANLWVPSTKCS
             274 0 STLLCEEGCAVVVDTGSSFISAPTSSLK
RENI$RAT      89 0 GIGTPSQTFKVIFDTGSANLWVPSTKCG
             274 0 ATLLCEEGCMAVVDTGTSYISGPTSSLQ
RENS$MOUSE    88 0 GIGTPPQTFKVIFDTGSANLWVPSTKCS
             273 0 STLLCEEGCEVVVDTGSSFISAPTSSLK
VPRT$AIDSS   175 2 TLWLDDKMFTGLIDTGADVTIIklEDWP
VPRT$HTLV2    41 0 VMGQTPQPTQALLDTGADLTVIPQTLVP
VPRT$MPMV    175 2 TLWLDDKMFTGLIDTGADVTIIklEDWP
Y5$CAMVC      32 6 FkGyKKiELhCfVDTGASLCIASKfVIP
Y5$CAMVD      34 6 FkGyKKiELhCfVDTGASLCIASKfVIP
Y5$CAMVS      32 6 FkGyKKiELhCfVDTGASLCIASKfVIP
Y5$CERV       21 4 PGYQTNlDLhCyVDTGSSLCMASKyVIP
```

Fig.1. (a) Property pattern for aspartic proteases. The exclamation point (!) marks positions where no deviation from the pattern is allowed. Properties are either present (1) in all amino acids or are forbidden (0) in a given position. (b) Amino acid allowed in the respective positions. (c) First column: SWISSPROT codes; second column: position in the sequence, third column: number of mismatches; fourth column: detected sequence sections (lower case letters stand for mismatches). In addition to known aspartic proteases with their two domains, HIV-proteases and a group of other proteins of viral origin (called protein 5, e.g. Y5$CERV) also match the pattern. No false (i.e. structurally or functionally unrelated) protein was recognized up to 6 mismatches.

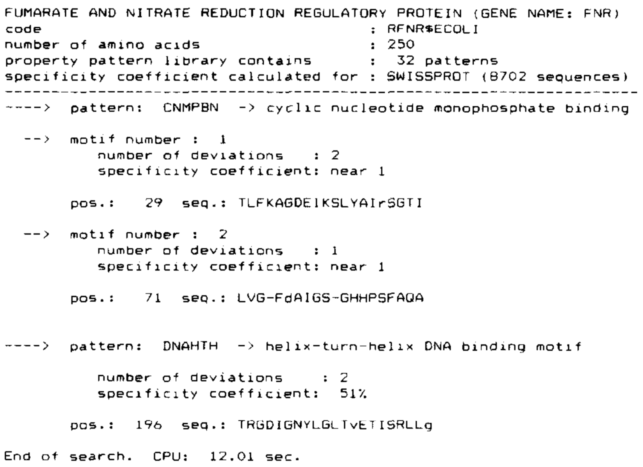


* "Y" means that structures of equivalent motifs are deposited in the Brookhaven Protein Data Bank [17]. By "Y" structural information is known from other sources.

** Only those references are cited, where the initial alignment was taken from. More detailed information about the patterns can be found therein.

Fig.2. Property pattern search in SWISSPROT (9.0, containing 8702 sequences). In the 32 patterns, 1532 motifs were correctly found (= = =) without any erroneous assignment. In addition to this finding, 134 motifs were correctly recognized (+ + +) with a loosened stringency criterion (empirical specificity coefficient greater than 60%, see text). With further 'loosening' of the patterns, more correct motifs may be found, but the number of false predictions increases rapidly.

library were: the kringle domain of clotting factors (present elsewhere as well, e.g. in apolipoprotein A), the GLA domain of vitamin-K-dependent proteins, two variants of the EGF domain with some differences in the cysteine patterns, Van Willebrand Factor A repeat (found also in other cell adhesion proteins), a pattern



```
FUMARATE AND NITRATE REDUCTION REGULATORY PROTEIN (GENE NAME: FNR)
code                                      : RFNR$ECOLI
number of amino acids                     : 250
property pattern library contains         :  32 patterns
specificity coefficient calculated for : SWISSPROT (8702 sequences)
-----------------------------------------------------------------------
---->  pattern:  CNMPBN  -> cyclic nucleotide monophosphate binding

  -->  motif number :  1
          number of deviations   : 2
          specificity coefficient: near 1

       pos.:   29  seq.: TLFKAGDEIKSLYAIrSGTI

  -->  motif number :  2
          number of deviations   : 1
          specificity coefficient: near 1

       pos.:   71  seq.: LVG-FdAIGS-GHHPSFAQA


---->  pattern:  DNAHTH  -> helix-turn-helix DNA binding motif

          number of deviations    : 2
          specificity coefficient:  51%

       pos.:  196  seq.: TRGDIGNYLGLTvETISRLLg

End of search.  CPU:  12.01 sec.
```

Fig.3. Analysis of fumarate and nitrate reduction regulatory protein from *E. coli* with screening against the property pattern library. Lower case letters in the printed sequence segment indicate deviations from the pattern.

common to low-density lipoprotein receptors and some complement components, a domain of many C3 binding proteins, two conserved regions of the anaphylotoxins, a repeat of the integrin β-chains (a cell adhesion receptor protein family), and the lectin domain.

The third group comprises regions localized in active centers of members of protein superfamilies. If such a pattern is detected in a sequence, then the whole topology of the protein can be predicted. Some different endoproteinases like aspartic proteases (see fig.1) of the pepsin family, serine proteases of the trypsin family, cysteine proteases of the papain type and those from some viruses were investigated. To our surprise, the proposed homology of serine proteases with some viral cysteine proteases [19,20] could not be found automatically with our method. In addition to the cystatin-kininogen family of cystein protease inhibitors, the serine protease inhibitors of the Bowman-Birk family, bovine pancreas trypsin family and the serpins were also included. Other specimens of this group are protein kinases (containing another ATP binding site as described above), helicases (also containing an ATP binding site) and ribonucleases (A and B).

To test the quality of the collected set of property patterns, we screened the 8702 entries of SWISSPROT database for the 32 patterns. With the appropriate choice of the stringency criteria we were able to detect 1532 sequence segments of 1065 proteins that correspond to one of the consensus patterns (fig.2). No segment with a known function or structure other than what was looked for was found. In 76 further cases, which were not directly confirmed by literature but by other circumstances (function, pathway, etc.) the assignment is likely to be correct. Loosening the matching stringency criterion by allowing more mismatches and setting the 'specificity coefficient' greater than 60%, 134 further motifs correctly appear (fig.2) as well as some 'false' predictions. In our opinion specificity is more important than exhaustive detection, and the empirical 'specificity coefficient' should help to estimate the reliability of prediction (fig.3).

In summary, we propose to make use of our (expanding) set of functional regions and structural motif as a heuristic tool for theoretical characterization of new proteins whose sequences have been derived, e.g. by genomic analysis (for example, see fig.3). The set of property patterns and the surrounding software running on VMS-compatible systems are available from the author on request.

Acknowledgements: I thank Professor J.G. Reich for helpful hints, discussions and critical reading of the manuscript. The work was supported by Professor W. Pfeil and Dr W. Schöpp.

REFERENCES

[1] Hodgman, T.C. (1986) Comp. Appl. Biosci. 2, 181-187.

[2] Patthy, L. (1988) J. Mol. Biol. 202, 680-696.

[3] Gribskov, M., Homyak, M., Edenfield, J. and Eisenberg, D. (1988) Comp. Appl. Biosci. 4, 61-66.

[4] Barker, W.C., Hunt, L.T. and George, D.G. (1988) Prot. Seq. Data Anal. 1, 363-373.

[5] Maulik, S. (1989) Prot. Seq. Data Anal. 2, 111-114.

[6] Abarbanel, R.A. (1984) Nucleic Acids Res. 12, 263-280.

[7] Taylor, W.R. (1986) J. Mol. Biol. 188, 233-258.

[8] Patthy, L. (1987) J. Mol. Biol. 198, 567-577.

[9] Gribskov, M., McLachlan, A.D. and Eisenberg, D. (1987) Proc. Natl. Acad. Sci. USA 84, 4355-4358.

[10] Staden, R. (1988) Comp. Appl. Biosci. 4, 53-60.

[11] Rooman, M.J. and Wodak, S.J. (1988) Nature 335, 45-49.

[12] Bork, P. and Grunwald, C. (1989) Stud. Biophys. 129, 231-241.

[13] Bork, P. and Grunwald, C. (1989) Eur. J. Biochem., submitted.

[14] Taylor, W.R. (1988) Prot. Eng. 2, 77-86.

[15] Pearl, L.H. and Taylor, W.R. (1987) Nature 329, 351-354.

[16] Miller, M., Jaskolski, M., Rao, J.K.M., Leis, J. and Wlodawer, A. (1989) Nature 337, 576-579.

[17] Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., jr., Brice, M.D., Rodgers, J.A., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) J. Mol. Biol. 112, 535-542.

[18] Patthy, L. (1987) FEBS Lett. 214, 1-7.

[19] Baza, J.F. and Fletterick, R.J. (1988) Proc. Natl. Acad. Sci. USA 85, 7872-7876.

[20] Gorbalenya, A.E., Donchenkov, A.P., Blinov, V.M. and Koonin, E.V. (1989) FEBS Lett. 243, 103-114.

[21] Kretsinger, R.H. (1987) Cold Spring Harb. Symp. Quant. Biol. 52, 499-510.

[22] Vyas, N.K., Vyas, M.N. and Quiocho, F.A. (1987) Nature 327, 635-637.

[23] Payre, F. and Vincent, A. (1988) FEBS Lett. 234, 245-250.

[24] Gibson, T.J., Postna, P.M., Brown, R.S. and Argos, P. (1988) Prot. Eng. 2, 209-218.

[25] Pabo, C.O. and Sauer, R.T. (1984) Annu. Rev. Biochem. 53, 291-321.

[26] Brennan, R.G. and Matthews, B.W. (1989) J. Biol. Chem. 264, 1903-1906.

[27] Dreyfuss, G., Swanson, M.S. and Pinol-Roma, S. (1988) Trends Biochem. Sci. 13, 86-91.

[28] Wierenga, R.K., Terpsta, P. and Hol, W.G.J. (1986) J. Mol. Biol. 187, 101-107.

[29] Rice, D.W., Schultz, G.E. and Guest, J.R. (1984) J. Mol. Biol. 174, 483-496.

[30] Walker, J.E., Saraste, M., Runswick, W.J. and Gay, N.J. (1982) EMBO J. 1, 945-951.

[31] Dever, T.E., Glynias, M.J. and Merrick, W.C. (1987) Proc. Natl. Acad. Sci. USA 84, 1814-1818.

[32] Möller, W. and Amons, R. (1985) FEBS Lett. 186, 1-7.

[33] Takio, K., Wade, R.D., Smith, S.B., Krebs, E.G., Walsh, K.A. and Titani, K. (1984) Biochemistry 23, 4201-4218.

[34] North, A.C.T. (1989) J. Biol. Macromol. 11, 56-58.

[35] Tulinski, A., Park, C.H. and Skrypczak-Jarkun, E. (1988) J. Mol. Biol. 202, 885-901.

[36] Doolittle, R.F., Feng, D.F. and Johnson, M.S. (1984) Science 231, 558-560.

[37] Reid, K.B.M., Bentley, D.R., Campbell, R.D., Chung, L.P., Sin, R.B., Kristensen, T. and Tack, B.F. (1986) Immunol. Today 7, 230-240.

[38] Pytela, R. (1988) EMBO J. 7. 1371-1378.

[39] Kishimoto, T.K., O'Conner, K., Lee, A., Roberts, T.M. and Springer, T.A. (1987) Cell 48, 681-690.

[40] Greer, J. (1986) Enzyme 36, 150-163.

[41] Peterson, T.E. (1988) FEBS Lett. 231, 51-53.

[42] Bairoch, R. and Clavery, J.M. (1988) Nature 329, 88.

[43] Barrett, A.J. (1986) in: Proteinase Inhibitors (Barrett, A.J. and Salvesen, G. eds) pp. 3-54, Elsevier, Amsterdam.

[44] Subramanian, E. (1978) in: Biomolecular Structure, Conformation, Function and Evolution, vol. 1 (Srinivasan, R., Subramanian, E. and Yathinda, S. eds) pp. 19-31, Pergamon Press, Oxford.

[45] Müller-Esterl, W., Fritz, W., Kellermann, J., Lottspeich, F., Machleidt, W. and Turk, V. (1985) FEBS Lett. 191, 221-226.

[46] Ikenaka, T. and Norioka, S. (1986) in: Proteinase Inhibitors (Barrett, A.J. and Salvesen, G. eds) pp. 361-374, Elsevier, Amsterdam.

[47] Creighton, T.E. and Charles, I.G. (1987) Cold Spring Harb. Symp. Quant. Biol. 52, 511-519.

[48] Carrell, R.W., Pemberton, P.A. and Boswell, D.R. (1987) Cold Spring Harb. Symp. Quant. Biol. 52, 527-535.

[49] Beintema, J.J., Schweller, C., Irie, M. and Carsana, A. (1988) Prog. Biophys. Mol. Biol. 51, 165-192.

[50] Hodgman, T.C. (1988) Nature 333, 22-23.

[51] Gorbalenya, A.E., Koonin, E.V., Donchenkov, A.P. and Blinov, V.M. (1988) Nature 333, 22.