



Article scientifique

Article

1993

Published version

Open Access

This is the published version of the publication, made available in accordance with the publisher's policy.

The PROSITE dictionary of sites and patterns in proteins, its current status

Bairoch, Amos Marc

How to cite

BAIROCH, Amos Marc. The PROSITE dictionary of sites and patterns in proteins, its current status. In: Nucleic acids research, 1993, vol. 21, n° 13, p. 3097–3103. doi: 10.1093/nar/21.13.3097

This publication URL: <https://archive-ouverte.unige.ch/unige:36853>

Publication DOI: [10.1093/nar/21.13.3097](https://doi.org/10.1093/nar/21.13.3097)

The PROSITE dictionary of sites and patterns in proteins, its current status

Amos Bairoch

Department of Medical Biochemistry, University of Geneva, 1 rue Michel Servet, 1211 Geneva 4, Switzerland

BACKGROUND

PROSITE is a compilation of sites and patterns found in protein sequences; it can be used as a method of determining the function of uncharacterized proteins translated from genomic or cDNA sequences. In some cases the sequence of an unknown protein is too distantly related to any protein of known structure to detect its resemblance by overall sequence alignment, but relationships can be revealed by the occurrence in its sequence of a particular cluster of residue types which is variously known as a pattern, motif, signature, or fingerprint. These motifs arise because specific region(s) of a protein which may be important, for example, for their binding properties or for their enzymatic activity are conserved in both structure and sequence. These structural requirements impose very tight constraints on the evolution of these small but important portion(s) of a protein sequence. The use of protein sequence patterns to determine the function of proteins is becoming very rapidly one of the essential tools of sequence analysis. This reality has been recognized by many authors [1,2]. While there have been a number of reviews of published patterns [3,4,5], no attempt had been made until very recently [6,7] to systematically collect biologically significant patterns or to discover new ones. Based on these observations, we decided in 1988, to actively pursue the development of a database of patterns which would be used to search against sequences of unknown function. This database, called PROSITE, contains some patterns which have been published in the literature, but the majority have been developed in the last four years by the author.

LEADING CONCEPTS

The design of PROSITE follows four leading concepts:

Completeness. For such a compilation to be helpful in the determination of protein function, it is important that it contains as many biologically meaningful patterns as possible.

High specificity of the patterns. In the majority of cases we have chosen patterns that are specific enough that they do not detect too many unrelated sequences, yet they will detect most, if not all, sequences that clearly belong to the set in consideration.

Documentation. Each of the patterns is fully documented; the documentation includes a concise description of the protein family that it is designed to detect as well as a summary of the reasons leading to the development of the pattern.

Periodic reviewing. It is important that each pattern be periodically reviewed to insure that it is still valid.

FORMAT

The PROSITE database is composed of two ASCII (text) files. The first file (PROSITE.DAT) is a computer-readable file that contains all the information necessary for programs that make use of PROSITE to scan sequence(s) for the occurrence of the patterns. This file also includes, for each of the patterns described, statistics on the number of hits obtained while scanning for that pattern in the SWISS-PROT protein sequence data bank [8]. Cross-references to the corresponding SWISS-PROT entries are also present in the file. The second file (PROSITE.DOC), which we call the textbook, contains textual information that documents each pattern. A user manual (PROSUSER.TXT) is distributed with the database; it fully describes the format of both files. A sample textbook entry is shown in Figure 1 with the corresponding data from the pattern file.

CONTENT OF THE CURRENT RELEASE

Release 10.1 of PROSITE (April 1993) contains 635 documentation entries describing 803 different patterns. The list of these entries is provided in Appendix 1. The database requires about 2 Mb of disk storage space.

COMPUTER PROGRAMS THAT MAKE USE OF PROSITE

Many academic groups and commercial companies have developed computer programs that make use of PROSITE. We list here some of these programs (a full descriptive list is included with the database and is stored in a file called 'PROSITE.PRG').

ACADEMIC

Program	Operating system	Author
MacPattern	Apple McIntosh	Rainer Fuchs [9]
prosite.c	IBM 3090-400E and Unix	Klaus Hartmuth
ProSearch	Unix and DOS (AWK)	Lee Kolakowski [10]
cregex	Unix and DOS	Jack Leunissen
dbsite/mksite	Unix	J.-M. Claverie
PROINDEX	VAX VMS	Steve Clark
Quelsite	VAX VMS	Claude Valencien
Scrutineer	VAX VMS and Unix	Peter Sibbald [11]
PATTERN	Unix	Olivier Boulot
PIP and PIPi	VAX VMS and Unix	Roger Staden [12]
PATMAT	OS and Unix	Steven Henikoff [13]
PROTOMAT	OS and Unix	Steven Henikoff [14]

COMMERCIAL

Program	Package	Supplier	Operating system
MOTIF	GCG	Genetics Comp. Group	Vax VMS and Unix
QUEST	IG-Suite	IntelliGenetics	Vax VMS and Unix
PROMOT		OML	Vax VMS and Unix
PROSITE	PC/Gene	IntelliGenetics	DOS
PROSITE	GeneWorks	IntelliGenetics	Apple McIntosh
Protean	LaserGene	DNASTAR	Apple McIntosh
PROTSITE		National Biosciences	DOS
SEQ/Pattern	ProExplore	Biostructure	Unix

FUTURE DEVELOPMENTS

There are a number of protein families as well as functional or structural domains that cannot be detected using patterns due to their extreme sequence divergence. Typical examples of important functional domains which are weakly conserved are the immunoglobulin domains, the SH2 and SH3 domains, and the fibronectin type III domain. In such domains there are only a few sequence positions which are well conserved. Any attempt to build a consensus pattern for such regions will either fail to pick up a significant proportion of the protein sequences that contain such a region (false negatives) or will pick up too many proteins that do not contain the region (false positives). Techniques based on the use of weight matrices [15,16,17] allows the detection of such proteins or domains. We plan to collaborate closely with Dr P. Bucher of the Swiss Cancer Research Institute (ISREC) in Lausanne and with Dr T.K. Attwood of the University of Leeds to develop such methods and to integrate them into PROSITE.

HOW TO OBTAIN PROSITE

PROSITE is distributed on magnetic tape and on CD-ROM by the EMBL Data Library. For all enquiries regarding the subscription and distribution of PROSITE one should contact:

EMBL Data Library
 European Molecular Biology Laboratory
 Postfach 10.2209, Meyerhofstrasse 1
 6900 Heidelberg, Germany
 Telephone: (+49 6221) 387 258
 Telefax: (+49 6221) 387 519 or 387 306
 Electronic network address: datalib@EMBL-heidelberg.de

PROSITE can be obtained from the EMBL File Server [18]. Detailed instructions on how to make the best use of this service, and in particular on how to obtain PROSITE, can be obtained by sending to the network address netserv@EMBL-heidelberg.de the following message:

HELP
 HELP PROSITE

If you have access to a computer system linked to the Internet you can obtain PROSITE using FTP (File Transfer Protocol), from the following file servers:

EMBL anonymous FTP server
 Internet address: [ftp.EMBL-heidelberg.de](ftp://EMBL-heidelberg.de) (or 192.54.41.33)

NCBI Repository (National Library of Medicine, NIH,
 Washington D.C., U.S.A.)
 Internet address: ncbi.nlm.nih.gov (130.14.20.1)

Basel Biozentrum Biocomputing server (EMBNET SWISS node)
 Internet address: bioftp.unibas.ch (or 131.152.8.1)

ExPASy (Expert Protein Analysis System server, University
 of Geneva, Switzerland)
 Internet address: expasy.hcuge.ch (129.195.254.61)

National Institute of Genetics (Japan) FTP server
 Internet address: [ftp.nig.ac.jp](ftp://nig.ac.jp) (133.39.16.66)

You can also browse through PROSITE using various Internet Gopher servers that specialize in biosciences (biogophers) [19]. Gopher is a distributed document delivery service that allows a neophyte user to access various types of data residing on multiple hosts in a seamless fashion.

The present distribution frequency is four releases per year. No restrictions are placed on use or redistribution of the data.

REFERENCES

- Doolittle R.F. (In) Of URFs and ORFs: a primer on how to analyze derived amino acid sequences., University Science Books, Mill Valley, California, (1986).
- Lesk A.M. (In) Computational Molecular Biology, Lesk A.M., Ed., pp17-26, Oxford University Press, Oxford (1988).
- Barker W.C., Hunt T.L., George D.G. *Protein Seq. Data Anal.* 1:363-373(1988).
- Hodgman T.C. *Comput. Appl. Biosci.* 5:1-13(1989).
- Taylor W.R., Jones D.T. *Curr. Opin. Struct. Biol.* 1:327-333(1991).
- Bork P. *FEBS Lett.* 257:191-195(1989).
- Smith H.O., Annau T.M., Chandrasegaran S. *Proc. Natl. Acad. Sci. USA* 87:826-830(1990).
- Bairoch A., Boeckmann B. *Nucleic Acids Res.* 20:2019-2022(1992).
- Fuchs R. *Comput. Appl. Biosci.* 7:105-106(1991).
- Kolakowski L.F. Jr., Leunissen J.A.M., Smith J.E. *Biotechniques* 13:919-921(1992).
- Sibbald P.R., Sommerfeldt H., Argos P. *Comput. Appl. Biosci.* 7:535-536(1991).
- Staden R. *DNA Sequence* 1:369-374(1991).
- Wallace J.C., Henikoff S. *Comput. Appl. Biosci.* 8:249-254(1992).
- Henikoff S., Henikoff J. *Nucleic Acids Res.* 19:6565-6572(1991).
- Staden R. *Nucleic Acids Res.* 12:505-519(1984).
- Gribskov M., Luethy R., Eisenberg D. *Meth. Enzymol.* 183:146-159(1990).
- Attwood T.K., Eliopoulos E.E., Findlay J.B.C. *Gene* 98:153-159(1991).
- Stoehr P.J., Omond R.A. *Nucleic Acids Res.* 17:6763-6764(1989).
- Gilbert D. *Trends Biochem. Sci.* 18:107-108(1993).

a

```
(PDOC00107)
(PS00116; DNA_POLYMERASE_B)
(BEGIN)
*****
* DNA polymerase family B signature *
*****
```

Replicative DNA polymerases (EC 2.7.7.7) are the key enzymes catalyzing the accurate replication of DNA. They require either a small RNA molecule or a protein as a primer for the de novo synthesis of a DNA chain. On the basis of sequence similarities a number of DNA polymerases have been grouped together [1 to 7] under the designation of DNA polymerase family B. The polymerases that belong to this family are:

- Higher eukaryotes polymerases alpha.
- Higher eukaryotes polymerases delta.
- Yeast polymerase I/alpha (gene POL1), polymerase II/epsilon (gene POL2), polymerase III/delta (gene POL3), and polymerase REV3.
- Escherichia coli polymerase I (gene dnaI or polB).
- Polymerases of viruses from the herpesviridae family.
- Polymerases from Adenoviruses.
- Polymerases from Baculoviruses.
- Polymerases from Chlorella viruses.
- Polymerases from Poxviruses.
- Bacteriophage T4 polymerase.
- Podoviridae bacteriophages Phi-29, M2, and PZA polymerase.
- Tectiviridae bacteriophage PRD1 polymerase.
- Polymerases encoded on eukaryotic linear DNA plasmids such as Kluyveromyces lactis pGKL1 and pGKL2, Agaricus bisporus pEM, Ascobolus immersus pA12, Claviceps purpurea pCLK1, and maize S-1.

Six regions of similarity (numbered from I to VI) are found in all or a subset of the above polymerases. The most conserved region (I) includes a perfectly conserved tetrapeptide which contains two aspartate residues. The function of this conserved region is not yet known, however it has been suggested [3] that it may be involved in binding a magnesium ion. We use this conserved region as a signature for this family of DNA polymerases.

- Consensus pattern: [YA]-[GLIVMSTAC]-D-T-D-[SG]-[LIVMFTC]-x-[LIVMSTAC]
- Sequences known to belong to this class detected by the pattern: ALL, except for yeast polymerase II/epsilon and Agaricus bisporus pEM.
- Other sequence(s) detected in SWISS-PROT: chicken vitellogenin 2.
- Last update: May 1993 / Pattern and text revised.

- [1] Jung G., Leavitt M.C., Maieh J.-C., Ito J.
Proc. Natl. Acad. Sci. U.S.A. 84:8287-8291(1987).
[2] Bernad A., Zaballos A., Salas M., Blanco L.
EMBO J. 6:4219-4225(1987).
[3] Argos P.
Nucleic Acids Res. 16:9909-9916(1988).
[4] Wang T.S.-F., Wong S.W., Korn D.
FASEB J. 3:14-21(1989).
[5] Delarue M., Poch O., Todro M., Moras D., Argos P.
Protein Engineering 3:461-467(1990).
[6] Ito J., Braithwaite D.K.
Nucleic Acids Res. 19:4045-4057(1991).
[7] Braithwaite D.K., Ito J.
Nucleic Acids Res. 21:787-802(1993).
(END)

b

```
ID DNA POLYMERASE_B; PATTERN.
AC PS00116;
DT APR-1990 (CREATED); MAY-1993 (DATA UPDATE); MAY-1993 (INFO UPDATE).
DE DNA polymerase family B signature.
PA [YA]-[GLIVMSTAC]-D-T-D-[SG]-[LIVMFTC]-x-[LIVMSTAC].
NR /RELEASE=24,28154;
NR /TOTAL=42(42); /POSITIVE=41(41); /UNKNOWN=0(0); /FALSE_POS=1(1);
NR /FALSE_NEG=2(2);
CC /TAXO_RANGE=ABEPV; /MAX_REPEAT=1;
DR P26019, DPOA_DROME, T; P09684, DPOA_HUMAN, T; P28040, DPOA_SCHPO, T;
DR P27727, DPOA_TRYBB, T; P13382, DPOA_YEAST, T; P28339, DPOD_BOVIN, T;
DR P28340, DPOD_HUMAN, T; P15436, DPOD_YEAST, T; P14284, DPOD_YEAST, T;
DR P21189, DPOZ_ECOLI, T; P26811, DPOL_SULSO, T; P03261, DPOL_ADE02, T;
DR P04495, DPOL_ADE05, T; P05664, DPOL_ADE07, T; P06538, DPOL_ADE12, T;
DR P03198, DPOL_EBV, T; P08546, DPOL_NCIVA, T; P04293, DPOL_HSV11, T;
DR P07917, DPOL_HSV1A, T; P04292, DPOL_HSV1K, T; P09854, DPOL_HSV1S, T;
DR P07918, DPOL_HSV21, T; P28857, DPOL_HSV6U, T; P28858, DPOL_HSV6B, T;
DR P28859, DPOL_HSV11, T; P24907, DPOL_HSV5A, T; P09252, DPOL_VZVO, T;
DR P09804, DPO1_KLULA, T; P05468, DPO2_KLULA, T; P22373, DPOL_CLAPU, T;
DR P10582, DPOL_MAIZE, T; P18131, DPOL_MPVAC, T; P20509, DPOL_VACCC, T;
DR P06856, DPOL_VACCV, T; P21402, DPOL_POMPV, T; P03680, DPOL_BPPH2, T;
DR P06950, DPOL_BPPZA, T; P19894, DPOL_BPH2, T; P10479, DPOL_BPPRD, T;
DR P04415, DPOL_BPT4, T;
DR P22374, DPOL_ASCIM, N; P21951, DPOE_YEAST, N;
DR P02845, VITZ_CHICK, F;
DO PDOC00107;
//
```

Figure 1. Sample data from PROSITE. a) A documentation (textbook) entry. b) The corresponding entry in the pattern file.

Appendix 1. List of patterns documentation entries in release 10.1 of PROSITE

Post-translational modifications

N-glycosylation site
Glycosaminoglycan attachment site
Tyrosine sulfatation site
cAMP- and cGMP-dependent protein kinase phosphorylation site
Protein kinase C phosphorylation site
Casein kinase II phosphorylation site
Tyrosine kinase phosphorylation site
N-myristoylation site
Amidation site
Aspartic acid and asparagine hydroxylation site
Vitamin K-dependent carboxylation domain
Phosphopantetheine attachment site
Prokaryotic membrane lipoprotein lipid attachment site
Prokaryotic N-terminal methylation site
Farnesyl group binding site (CAAX box)

Domains

Endoplasmic reticulum targeting sequence
Microbodies C-terminal targeting signal
Gram-positive cocci surface proteins 'anchoring' hexapeptide
Bipartite nuclear targeting sequence
Cell attachment sequence
ATP/GTP-binding site motif A (P-loop)
EF-hand calcium-binding domain
Actinin-type actin-binding domain signatures
Cofilin/tropomyosin-type actin-binding domain
Apple domain
Band 4.1 family domain signatures
Kringle domain signature

EGF-like domain cysteine pattern signature
Fibrinogen beta and gamma chains C-terminal domain signature
Type II fibronectin collagen-binding domain
Hemopexin domain signature
C-type lectin domain signature
Osteonectin domain signatures
Somatomedin B domain signature
Thyroglobulin type-1 repeat signature
'Trefoil' domain signature
Cellulose-binding domain, bacterial type
Cellulose-binding domain, fungal type
Chitin recognition or binding domain signature
Barwin domain signatures
WAP-type 'four-disulfide core' domain signature
Phorbol esters/diacylglycerol binding domain
C2 domain signature
ZP domain signature

DNA or RNA associated proteins

'Homeobox' domain signature
'Homeobox' antennapedia-type protein signature
'Homeobox' engrailed-type protein signature
'Paired box' domain signature
'POU' domain signatures
Zinc finger, C2H2 type, domain
Zinc finger, C³HC4 type, signature
Nuclear hormones receptors DNA-binding region signature
GATA-type zinc finger domain
Poly(ADP-ribose) polymerase zinc finger domain
Fungal Zn(2)-Cys(6) binuclear cluster domain

Appendix 1. *continued*

- Leucine zipper pattern
 Fos/jun DNA-binding basic domain signature
 Myb DNA-binding domain repeat signatures
 Myc-type, 'helix-loop-helix' putative DNA-binding domain signature
 p53 tumor antigen signature
 CBF/NF-Y subunits signatures
 'Cold-shock' DNA-binding domain signature
 CTF/NF-I signature
 Ets-domain signatures
 Fork head domain signatures
 HSF-type DNA-binding domain signature
 IRF family signature
 LIM domain
 SRF-type transcription factors DNA-binding and dimerization domain
 TEA domain signature
 Transcription factor TFIIB repeat signature
 Transcription factor TFIID repeat signature
 TFIIS cysteine-rich domain signature
 DEAD and DEAH box families ATP-dependent helicases signature
 Eukaryotic putative RNA-binding region RNP-1 signature
 Fibrillarin signature
 XPAC protein signatures
 Bacterial regulatory proteins, araC family signature
 Bacterial regulatory proteins, asnC family signature
 Bacterial regulatory proteins, crp family signature
 Bacterial regulatory proteins, gntR family signature
 Bacterial regulatory proteins, lacI family signature
 Bacterial regulatory proteins, luxR family signature
 Bacterial regulatory proteins, lysR family signature
 Bacterial regulatory proteins, merR family signature
 Transcriptional antiterminators bglG family signature
 Sigma-54 factors family signatures
 Sigma-70 factors family signatures
 Sigma-54 interaction domain signatures
 Single-strand binding protein family signatures
 Bacterial histone-like DNA-binding proteins signature
 Histone H2A signature
 Histone H2B signature
 Histone H3 signature
 Histone H4 signature
 HMG1/2 signature
 HMG-I and HMG-Y DNA-binding domain (A T-hook)
 HMG14 and HMG17 signature
 Bromodomain
 Chromo domain
 Regulator of chromosome condensation signatures
 Protamine P1 signature
 Nuclear transition protein 1 signature
 Ribosomal protein L2 signature
 Ribosomal protein L3 signature
 Ribosomal protein L5 signature
 Ribosomal protein L6 signatures
 Ribosomal protein L9 signature
 Ribosomal protein L11 signature
 Ribosomal protein L13 signature
 Ribosomal protein L14 signature
 Ribosomal protein L15 signature
 Ribosomal protein L16 signatures
 Ribosomal protein L22 signature
 Ribosomal protein L23 signature
 Ribosomal protein L29 signature
 Ribosomal protein L30 signature
 Ribosomal protein L33 signature
 Ribosomal protein L34 signature
 Ribosomal protein L19e signature
 Ribosomal protein L30e signature
 Ribosomal protein L32e signature
 Ribosomal protein L46e signature
 Ribosomal protein S3 signatures
 Ribosomal protein S4 signature
 Ribosomal protein S5 signature
 Ribosomal protein S7 signature
 Ribosomal protein S8 signature
 Ribosomal protein S9 signature
 Ribosomal protein S10 signature
 Ribosomal protein S11 signature
 Ribosomal protein S12 signature
 Ribosomal protein S13 signature
 Ribosomal protein S14 signature
 Ribosomal protein S15 signature
 Ribosomal protein S16 signature
 Ribosomal protein S17 signature
 Ribosomal protein S18 signature
 Ribosomal protein S19 signature
 Ribosomal protein S4e signature
 Ribosomal protein S6e signature
 Ribosomal protein S17e signature
 Ribosomal protein S19e signature
 Ribosomal protein S24e signature
 Ribosomal protein S26e signature
 DNA mismatch repair proteins mutL/hexB/PMS1 signature
 DNA mismatch repair proteins mutS family signature
 RecF protein signatures
 Small, acid-soluble spore proteins, alpha/beta type, signatures
- Enzymes**
- Oxidoreductases**
- Zinc-containing alcohol dehydrogenases signature
 Iron-containing alcohol dehydrogenases signature
 Short-chain alcohol dehydrogenase family signature
 Aldo/keto reductase family signatures
 Histidinol dehydrogenase active site
 L-lactate dehydrogenase active site
 D-isomer specific 2-hydroxyacid dehydrogenases signatures
 Hydroxymethylglutaryl-coenzyme A reductases signatures
 3-hydroxyacyl-CoA dehydrogenase signature
 Malate dehydrogenase active site signature
 Malic enzymes signature
 Isocitrate and isopropylmalate dehydrogenases signature
 6-phosphogluconate dehydrogenase signature
 Glucose-6-phosphate dehydrogenase active site
 IMP dehydrogenase/GMP reductase signature
 Bacterial quinoprotein dehydrogenases signatures
 FMN-dependent alpha-hydroxy acid dehydrogenases active site
 GMC oxidoreductases signatures
 Eukaryotic molybdopterin oxidoreductases signature
 Prokaryotic molybdopterin oxidoreductases signatures
 Aldehyde dehydrogenases active sites
 Glyceraldehyde 3-phosphate dehydrogenase active site
 Fumarate reductase/succinate dehydrogenase FAD-binding site
 Acyl-CoA dehydrogenases signatures
 Glutamate/Leucine/Phenylalanine dehydrogenases active site
 D-amino acid oxidase signature
 Delta 1-pyrroline-5-carboxylate reductase signature
 Dihydrofolate reductase signature
 Tetrahydrofolate dehydrogenase/cyclohydrolase signatures
 Pyridine nucleotide-disulphide oxidoreductases class-I active site
 Pyridine nucleotide-disulphide oxidoreductases class-II active site
 Respiratory-chain NADH dehydrogenase subunit 1 signatures
 Respiratory-chain NADH dehydrogenase 30 Kd subunit signature
 Respiratory-chain NADH dehydrogenase 49 Kd subunit signature
 Respiratory-chain NADH dehydrogenase 51 Kd subunit signatures
 Respiratory-chain NADH dehydrogenase 75 Kd subunit signatures
 Nitrite reductases and sulfite reductase putative siroheme-binding sites
 Uricase signature
 Cytochrome c oxidase subunit I, copper B binding region signature
 Cytochrome c oxidase subunit II, copper A binding region signature
 Multicopper oxidases signatures
 Peroxidases signatures
 Catalase signatures
 Glutathione peroxidases signatures
 Lipoygenases, putative iron-binding region signatures
 Extradiol ring-cleavage dioxygenases signature
 Intradiol ring-cleavage dioxygenases signature
 Bacterial ring hydroxylating dioxygenases alpha-subunit signature
 Bacterial luciferase subunits signature
 Biotpterin-dependent aromatic amino acid hydroxylases signature
 Copper type II, ascorbate-dependent monooxygenases signatures
 Tyrosinase signatures
 Fatty acid desaturases signatures

Cytochrome P450 cysteine heme-iron ligand signature
 Heme oxygenase signature
 Copper/Zinc superoxide dismutase signatures
 Manganese and iron superoxide dismutases signature
 Ribonucleotide reductase large subunit signature
 Ribonucleotide reductase small subunit signature
 Nitrogenases component 1 alpha and beta subunits signatures
 NifH/frxC family signatures
 Nickel-dependent hydrogenases large subunit signatures
 Glutamyl-tRNA reductase signature

Transferases

Thymidylate synthase active site
 Methylated-DNA-protein-cysteine methyltransferase active site
 N-6 Adenine-specific DNA methylases signature
 N-4 cytosine-specific DNA methylases signature
 C-5 cytosine-specific DNA methylases signatures
 Serine hydroxymethyltransferase pyridoxal-phosphate attachment site
 Phosphoribosylglycinamide formyltransferase active site
 Aspartate and ornithine carbamoyltransferases signature
 Transketolase signatures
 Acyltransferases ChoActase/COT/CPT-II family signatures
 Thiolases signatures
 Chloramphenicol acetyltransferase active site
 cysE/lacA/nodL acetyltransferases signature
 Beta-ketoacyl synthases active site
 Chalcone and stilbene synthases active site
 Gamma-glutamyltranspeptidase signature
 Transglutaminases active site
 Phosphorylase pyridoxal-phosphate attachment site
 UDP-glucuronosyl and UDP-glucosyl transferases signature
 Purine/pyrimidine phosphoribosyl transferases signature
 Glutamine amidotransferases class-I active site
 Glutamine amidotransferases class-II active site
 Thymidine phosphorylase signature
 S-Adenosylmethionine synthetase signatures
 Polyprenyl synthetases signatures
 Riboflavin synthase alpha chain family Lum-binding site signature
 Dihydropteroate synthase signatures
 EPSP synthase active site
 Aspartate aminotransferases pyridoxal-phosphate attachment site
 Aminotransferases class-II pyridoxal-phosphate attachment site
 Aminotransferases class-III pyridoxal-phosphate attachment site
 Aminotransferases class-IV signature
 Phosphoserine aminotransferase signature
 Hexokinases signature
 Galactokinase signature
 GHMP kinases putative ATP-binding domain
 Phosphofructokinase signature
 pfkB family of carbohydrate kinases signatures
 Phosphoribulokinase signature
 Thymidine kinase cellular-type signature
 Prokaryotic carbohydrate kinases signature
 Protein kinases signatures
 Pyruvate kinase active site signature
 Phosphoglycerate kinase signature
 Aspartokinase signature
 ATP:guanido phosphotransferases active site
 PTS Hpr component phosphorylation sites signatures
 PTS permeases phosphorylation sites signatures
 Adenylate kinase signature
 Nucleoside diphosphate kinases active site
 Phosphoribosyl pyrophosphate synthetase signature
 7,8-dihydro-6-hydroxymethylpterin-pyrophosphokinase signature
 Bacteriophage-type RNA polymerase family active site signature
 Eukaryotic RNA polymerase II heptapeptide repeat
 Eukaryotic RNA polymerases 30 to 40 Kd subunits signature
 DNA polymerase family A signature
 DNA polymerase family B signature
 DNA polymerase family X signature
 Galactose-1-phosphate uridyl transferase active site signature
 CDP-alcohol phosphatidyltransferases signature
 PEP-utilizing enzymes signatures
 Rhodanese signatures

Hydrolases

Phospholipase A2 active sites signatures
 Lipases, serine active site

Colipase signature
 Carboxylesterases type-B active site
 Pectinesterase signatures
 Alkaline phosphatase active site
 Histidine acid phosphatases signatures
 5'-nucleotidase signatures
 Fructose-1-6-bisphosphatase active site
 Serine/threonine specific protein phosphatases signature
 Tyrosine specific protein phosphatases active site
 Inositol monophosphatase family signatures
 Prokaryotic zinc-dependent phospholipase C signature
 3'5'-cyclic nucleotide phosphodiesterases signature
 cAMP phosphodiesterases class-II signature
 Sulfatases signatures
 AP endonucleases family 1 signatures
 AP endonucleases family 2 signatures
 Endonuclease III iron-sulfur binding region signature
 Ribonuclease III family signature
 Bacterial Ribonuclease P protein component signature
 Ribonuclease T2 family histidine active sites
 Pancreatic ribonuclease family signature
 Beta-amylase signatures
 Polygalacturonase active site
 Clostridium cellulases repeated domain signature
 Chitinases class I signatures
 Alpha-lactalbumin/lysozyme C signature
 Alpha-galactosidase signature
 Alpha-L-fucosidase putative active site
 Glycosyl hydrolases family 1 signatures
 Glycosyl hydrolases family 2 signatures
 Glycosyl hydrolases family 3 active site
 Glycosyl hydrolases family 5 signature
 Glycosyl hydrolases family 6 signatures
 Glycosyl hydrolases family 9 active sites signatures
 Glycosyl hydrolases family 10 active site
 Glycosyl hydrolases family 11 active site signatures
 Glycosyl hydrolases family 17 signature
 Glycosyl hydrolases family 31 signatures
 Glycosyl hydrolases family 32 active site
 Alkylbase DNA glycosidases alka family signature
 Uracil-DNA glycosylase signature
 S-adenosyl-L-homocysteine hydrolase signatures
 Cytosol aminopeptidase signature
 Aminopeptidase P and proline dipeptidase signature
 Methionine aminopeptidase signature
 Serine carboxypeptidases, active sites
 Zinc carboxypeptidases, zinc-binding regions signatures
 Serine proteases, trypsin family, active sites
 Serine proteases, subtilase family, active sites
 Serine proteases, V8 family, active sites
 Prolyl endopeptidase family serine active site
 ClpP proteases active sites
 Eukaryotic thiol (cysteine) proteases active sites
 Ubiquitin carboxyl-terminal hydrolase, putative active-site signature
 Eukaryotic aspartyl proteases active site
 Neutral zinc metallopeptidases, zinc-binding region signature
 Matrixins cysteine switch
 Insulinase family signature
 recA signature
 Proteasome subunits signature
 Signal peptidases I signatures
 Amidases signature
 Asparaginase/glutaminase active site
 Urease active site
 ArgE/dapE/CPG2 family signatures
 Dihydroorotase signatures
 Beta-lactamases classes -A, -C, and -D active site
 Beta-lactamases class B signatures
 Arginase and agmatinase signatures
 Adenosine and AMP deaminase signature
 Inorganic pyrophosphatase signature
 Acylphosphatase signatures
 ATP synthase alpha and beta subunits signature
 ATP synthase gamma subunit signature
 ATP synthase delta (OSCP) subunit signature
 ATP synthase a subunit signature
 ATP synthase c subunit signature

Appendix 1. continued

E1-E2 ATPases phosphorylation site
Sodium and potassium ATPases beta subunits signatures
Cutinase, serine active site

Lyases

DDC/GAD/HDC pyridoxal-phosphate attachment site
Prokaryotic ornithine and lysine decarboxylases pyridoxal-phosphate attachment site
Orotidine 5'-phosphate decarboxylase active site
Phosphoenolpyruvate carboxylase active sites
Phosphoenolpyruvate carboxykinase (GTP) signature
Phosphoenolpyruvate carboxykinase (ATP) signature
Indole-3-glycerol phosphate synthase signature
Ribulose biphosphate carboxylase large chain active site
Fructose-biphosphate aldolase class-I active site
Fructose-biphosphate aldolase class-II signatures
Malate synthase signature
Citrate synthase signature
KDPG and KHG aldolases active site signatures
Isocitrate lyase signature
DNA photolyases signatures
Eukaryotic-type carbonic anhydrases signature
Prokaryotic-type carbonic anhydrases signatures
Fumarate lyases signature
Aconitase family signature
Enolase signature
Serine/threonine dehydratases pyridoxal-phosphate attachment site
Enoyl-CoA hydratase/isomerase signature
Tryptophan synthase alpha chain signature
Tryptophan synthase beta chain pyridoxal-phosphate attachment site
Delta-aminolevulinic acid dehydratase active site
Dihydrodipicolinate synthetase signatures
Phenylalanine and histidine ammonia-lyases signature
Porphobilinogen deaminase cofactor-binding site
Guanylate cyclases signature
Chorismate synthase signatures
Ferrochelatase signature

Isomerases

Alanine racemase pyridoxal-phosphate attachment site
Aldose 1-epimerase putative active site
Cyclophilin-type peptidyl-prolyl cis-trans isomerase signature
FKBP-type peptidyl-prolyl cis-trans isomerase signatures
Triosephosphate isomerase active site
Xylose isomerase signatures
Phosphoglucose isomerase signatures
Phosphoglycerate mutase family phosphohistidine signature
Phosphoglucomutase & phosphomannomutase phosphohistidine signature
Methylmalonyl-CoA mutase signature
Eukaryotic DNA topoisomerase I active site
Prokaryotic DNA topoisomerase I active site
DNA topoisomerase II signature

Ligases

Aminoacyl-transfer RNA synthetases class-I signature
Aminoacyl-transfer RNA synthetases class-II signatures
WHEP-TRS domain signature
ATP-citrate lyase and succinyl-CoA ligases active site
Glutamine synthetase signatures
Ubiquitin-activating enzyme signature
Ubiquitin-conjugating enzymes active site
Formate-tetrahydrofolate ligase signatures
Adenylosuccinate synthetase active site
Argininosuccinate synthase signatures
Phosphoribosylglycinamide synthetase signature
ATP-dependent DNA ligase signatures

Others

sopenicillin N synthetase signatures
Site-specific recombinases signatures
Thiamine pyrophosphate enzymes signature
Biotin-requiring enzymes attachment site
2-oxo acid dehydrogenases acyltransferase component lipoyl binding site
Putative AMP-binding domain signature

Electron transport proteins

Cytochrome c family heme-binding site signature
Cytochrome b5 family, heme-binding domain signature

Cytochrome b/b6 signatures
Cytochrome b559 subunits heme-binding site signature
Thioredoxin family active site
Glutaredoxin active site
Type-1 copper (blue) proteins signature
2Fe-2S ferredoxins, iron-sulfur binding region signature
4Fe-4S ferredoxins, iron-sulfur binding region signature
High potential iron-sulfur proteins signature
Rieske iron-sulfur protein signatures
Flavodoxin signature
Rubredoxin signature
Electron transfer flavoprotein alpha-subunit signature

Other transport proteins

Class I metallothioneins signature
Ferritin iron-binding regions signatures
Bacterioferritin signature
Transferrins signatures
Plant hemoglobins signature
Hemerythrins signature
Arthropod hemocyanins/insect LSPs signatures
ATP-binding proteins 'active transport' family signature
Binding-protein-dependent transport systems inner membrane component signature
Serum albumin family signature
Transthyretin signatures
Avidin/Streptavidin family signature
Eukaryotic cobalamin-binding proteins signature
Lipocalin signature
Cytosolic fatty-acid binding proteins signature
LBP/BPI/CETP family signature
Plant lipid transfer proteins signature
Uteroglobin family signatures
Mitochondrial energy transfer proteins signature
Sugar transport proteins signatures
Sodium symporters signatures
Prokaryotic sulfate- and thiosulfate-binding proteins signatures
Amino acid permeases signature
Aromatic amino acids permeases signature
Anion exchangers family signatures
Bacterial protein export pilT protein family signature
GltP/dctA family of transporters signatures
MIP family signature
Neurotransmitters transporters signatures
General diffusion gram-negative porins signature
Eukaryotic porin signature
Insulin-like growth factor binding proteins signature

Structural proteins

43 Kd postsynaptic protein signature
Actins signatures
Annexins repeated domain signature
Clathrin light chains signatures
Clusterin signatures
Connexins signatures
Crystallins beta and gamma 'Greek key' motif signature
Dynamin family signature
Intermediate filaments signature
Involucrin signature
Kinesin motor domain signature
Myelin basic protein signature
Myelin P0 protein signature
Myelin proteolipid protein signature
Neuromodulin (GAP-43) signatures
Profilin signature
Surfactant associated polypeptide SP-C palmitoylation sites
Synapsins signatures
Synaptobrevin signature
Synaptophysin/synaptoporin signature
Tropomyosins signature
Tubulin subunits alpha, beta, and gamma signature
Tubulin-beta mRNA autoregulation signal
Tau and MAP proteins repeated region signature
Neuraxin and MAP1B proteins repeated region signature
F-actin capping protein alpha subunit signatures
F-actin capping protein beta subunit signature
Vinculin family signatures

Amyloidogenic glycoprotein signatures
 Cadherins extracellular repeated domain signature
 Insect flexible cuticle proteins signature
 Gas vesicles protein GVPa signatures
 Gas vesicles protein GVPc repeated domain signature
 Flagella basal body rod proteins signature
 Plant viruses icosahedral capsid proteins 'S' region signature
 Potexviruses and carlaviruses coat protein signature

Receptors

Neurotransmitter-gated ion-channels signature
 G-protein coupled receptors signature
 G-protein coupled receptors family 2 signatures
 Visual pigments (opsins) retinal binding site
 Bacterial rhodopsins retinal binding site
 Receptor tyrosine kinase class II signature
 Receptor tyrosine kinase class III signature
 Receptor tyrosine kinase class V signatures
 Growth factor and cytokines receptors family signatures
 TNFR/NGFR family cysteine-rich region signature
 Integrins alpha chain signature
 Integrins beta chain cysteine-rich domain signature
 Natriuretic peptides receptors signature
 Photosynthetic reaction center proteins signature
 Photosystem I psaA and psaB proteins signature
 Phytochrome chromophore attachment site
 Speract receptor repeated domain signature
 TonB-dependent receptor proteins signature
 Type-II membrane antigens family signature
 Bacterial chemotaxis sensory transducers signature

Cytokines and growth factors

Granulins signature
 HBGF/FGF family signature
 PTN/MK heparin-binding protein family signatures
 Nerve growth factor family signature
 Platelet-derived growth factor (PDGF) family signature
 Small cytokines (intercrine/chemokine) signatures
 TGF-beta family signature
 TNF family signature
 Wnt-1 family signature
 Interferon alpha and beta family signature
 Granulocyte-macrophage colony-stimulating factor signature
 Interleukin-1 signature
 Interleukin-2 signature
 Interleukin-6/G-CSF/MGF family signature
 Interleukin-7 signature
 Interleukin-10 signature
 LIF/OSM family signature

Hormones and active peptides

Adipokinetic hormone family signature
 Bombesin-like peptides family signature
 Calcitonin/CGRP/LAPP family signature
 Corticotropin-releasing factor family signature
 Granins signatures
 Gastrin/cholecystokinin family signature
 Glucagon/GIP/secretin/VIP family signature
 Glycoprotein hormones alpha chain signatures
 Glycoprotein hormones beta chain signatures
 Gonadotropin-releasing hormones signature
 Insulin family signature
 Natriuretic peptides signature
 Neurohypophysial hormones signature
 Pancreatic hormone family signature
 Parathyroid hormone family signature
 Pyrokinins signature
 Somatotropin, prolactin and related hormones signatures
 Tachykinin family signature
 Thymosin beta-4 family signature
 Cecropin family signature
 Mammalian defensins signature
 Insect defensins signature
 Endothelins/sarafotoxins signature

Toxins

Plant thionins signature
 Snake toxins signature

Myotoxins signature
 Heat-stable enterotoxins signature
 Aerolysin type toxins signature
 Shiga/ricin ribosomal inactivating toxins active site signature
 Channel forming colicins signature
 Hok/gef family cell toxic proteins signature
 Staphylococcal enterotoxins/Streptococcal pyrogenic exotoxins signatures
 Thiol-activated cytolytins signature
 Membrane attack complex components/perforin signature

Inhibitors

Pancreatic trypsin inhibitor (Kunitz) family signature
 Bowman-Birk serine protease inhibitors family signature
 Kazal serine protease inhibitors family signature
 Soybean trypsin inhibitor (Kunitz) protease inhibitors family signature
 Serpins signature
 Potato inhibitor I family signature
 Squash family of serine protease inhibitors signature
 Cysteine proteases inhibitors signature
 Tissue inhibitors of metalloproteinases signature
 Cereal trypsin/alpha-amylase inhibitors family signature
 Alpha-2-macroglobulin family thiolester region signature
 Disintegrins signature
 Lambdoid phages regulatory protein CIII signature

Others

Pentaxin family signature
 Immunoglobulins and major histocompatibility complex proteins signature
 Gram-negative pili assembly chaperone signature
 Prion protein signatures
 Cyclins signature
 Proliferating cell nuclear antigen signature
 Arrestins signature
 Chaperonins cpn60 signature
 Chaperonins cpn10 signature
 Chaperonins TCP-1 signatures
 Heat shock hsp70 proteins family signatures
 Heat shock hsp90 proteins family signature
 DnaJ domains signatures
 Protein secY signatures
 CDC48/PAS1/SEC18 family signature
 Ubiquitin signature
 Beta-transducin family Trp-Asp repeats signature
 Ras GTPase-activating proteins signature
 Guanine-nucleotide dissociation stimulators CDC25 family signature
 Guanine-nucleotide dissociation stimulators CDC24 family signature
 Stathmin family signature
 SRP54-type proteins GTP-binding domain signature
 GTP-binding elongation factors signature
 Eukaryotic initiation factor 5A hypusine signature
 Prokaryotic-type peptide chain release factors signature
 Calreticulin family signatures
 S-100/ICaBP type calcium binding protein signature
 Hemolysin-type putative calcium-binding region signature
 HlyD family secretion proteins signature
 P-II protein signatures
 14-3-3 proteins signatures
 Caseins alpha/beta signature
 Legume lectins signatures
 Vertebrate galactoside-binding lectin signature
 Lysosome-associated membrane glycoproteins signatures
 Glycophorin A signature
 Seminal vesicle protein I repeats signature
 Seminal vesicle protein II repeats signature
 Stress-induced proteins SRP1/TIP1 family signature
 Tissue factor signature
 HCP repeats signature
 Bacterial ice-nucleation proteins octamer repeat
 Cell cycle proteins ftsW/rodA/spoVE signature
 Enterobacterial virulence outer membrane protein signatures
 Staphylocoagulase repeat signature
 11-S plant seed storage proteins signature
 Dehydrins signature
 Germin family signature
 Small hydrophilic plant seed proteins signature
 Pathogenesis-related proteins BetvI family signature
 Thaumatin family signature