Abstract

PROSITE is a compilation of sites and patterns found in protein sequences. The use of protein sequence patterns (or motifs) 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. While there have been a number of recent reports that review published patterns, no attempt had been made until very recently \[5,6\] to systematically collect biologically significant patterns or to discover new ones. It is for these reasons that we have developed, since 1988, a dictionary of sites and patterns which we call PROSITE.
PROSITE: a dictionary of sites and patterns in proteins

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. The use of protein sequence patterns (or motifs) 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 recent reports [3,4] that review published patterns, no attempt had been made until very recently [5,6] to systematically collect biologically significant patterns or to discover new ones. It is for these reasons that we have developed, since 1988, a dictionary of sites and patterns which we call PROSITE.

Some of the patterns compiled in PROSITE have been published in the literature, but the majority have been developed, in the last two years, by the author.

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 to programs that make use of PROSITE to scan sequence(s) with pattern(s). 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 [7]. Cross-references to the corresponding SWISS-PROT entries are also present in that 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.

Leading concepts

The design of PROSITE follows four leading concepts:

1. Completeness. For such a compilation to be helpful in the determination of protein function, it is important that it contains a significant number of biologically meaningful patterns.
2. High specificity. In the majority of cases we have chosen patterns that are specific enough not to detect too many unrelated sequences, yet that detect most if not all sequences that clearly belong to the set in consideration.
3. Documentation. Each of the patterns is fully documented; the documentation includes a concise description of the family of protein that it is supposed to detect as well as an explanation on the reasons that led to the selection of the particular pattern.
4. Periodic reviewing. It is important that each pattern be periodically reviewed, so as to insure that it is still relevant.

Content of the current release

Release 6.10 of PROSITE (February 1991) contains 375 documentation entries describing 433 different patterns. The list of these entries is provided in Appendix 1.

Distribution

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: DATA LIB@EMBL.BITNET

PROSITE can be obtained from the EMBL File Server [8]. Detailed instructions on how to make best use of this service, and in particular on how to obtain PROSITE, can be obtained by sending to the network address NETSERV@EMBL.BITNET

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:

GenBank On-line Service [9]
Internet address: genbank.bio.net (134.172.1.160)

NCBI
Internet address: ncbi.nlm.nih.gov (130.14.20.1)

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

REFERENCES

1a) A documentation (textbook) entry

(PDOC00107)
(PSD0116; 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 5] under the designation of DNA polymerase family B. The polymerases that belong to this family are:

- Human polymerase alpha.
- Yeast polymerase I, polymerase III, and polymerase REV3.
- Polymerases of viruses from the herpesviridae family (Herpes type I and II, Epstein-Barr, Cytomegalovirus, and Varicella-zoster).
- Polymerases from Adenoviruses.
- Polymerases from Baculoviruses.
- Vaccinia virus polymerase.
- Bacteriophage T4 polymerase.
- Podoviridae bacteriophages Phi-29, M2, and P2A polymerase.
- Tectiviridae bacteriophage JD1 polymerase.
- Putative polymerases from yeast E. lactis linear plasmids pKL1 and pKL2.
- Putative polymerase from the maize mitochondrial plasmid-like 51 DNA.

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]-x-D-T-D-[LIVMT]
- Sequences known to belong to this class detected by the pattern: ALL.
- Other sequence(s) detected in SWISS-PROT: chicken vitellogenin 2.
- Note: the residue in position 1 is Tyr in every family B polymerases, except in phage T4 where it is Ala.

(END)

1b) The corresponding entry in the pattern file

ID DNA_POLYMERASE_B; PATTERN.
AC PS00116;
DT APR-1990 (CREATED); NOV-1990 (DATA UPDATE); NOV-1990 (INFO UPDATE).
DE DNA polymerase family B signature.
PA [YA]-x-D-T-D-[LIVMT].
NR /RELEASE=16,1836;
NR /TOTAL=26(26); /POSITIVE=25(25); /UNKNOWN=0(0); /FALSE_POS=1(1); /FALSE_NEG=0(0);
CC /TAXO-RANGE=78789; /MAX-REPEAT=1.
DR P00884, DPOASMBUIMAN, T; P13382, DPOASYEAST, T; P15436, DPOASYEAST, T;
DP P12824, DPOASBUN, T; P03261, DPOASDEOD, T; P04495, DPOASDEOD, T;
DR P05646, DPOASDEOD, T; P05538, DPOASDEOD, T; P04415, DPOASDEOD, T;
DR P05646, DPOASDEOD, T; P05538, DPOASDEOD, T; P04415, DPOASDEOD, T;
DR P08546, DPOASHEMVIA, T; P05198, DPOASBEV, T; P04295, DPOASBEV, T;
DR P07917, DPOASHEMVIA, T; P02922, DPOASHEMVIA, T; P00854, DPOASHEMVIA, T;
DR P07917, DPOASHEMVIA, T; P02922, DPOASHEMVIA, T; P00854, DPOASHEMVIA, T;
DR P07917, DPOASHEMVIA, T; P02922, DPOASHEMVIA, T; P00854, DPOASHEMVIA, T;
DR P07917, DPOASHEMVIA, T; P02922, DPOASHEMVIA, T; P00854, DPOASHEMVIA, T;
DR P05468, DPOASKLULA, T; P10582, DPOASMAI, T; P18131, DPOASMAI, T;
DR P06566, DPOASVACC, T; P03680, DPOASVACC, T; P06950, DPOASVACC, T;
DR P10479, DPOASVPRD, T;
DR P02845, VIT28CHICK, T;
DO POC00107;

Figure 1. Sample data from PROSITE
Appendix 1. List of patterns documentation entries in release 6.10 of PROSITE

Post-translational modifications
N-glycosylation site
Glycosaminoglycan attachment site
Tyrosine sulfation 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
Bacterial activator proteins, Leucine zipper
Poly(ADP-ribose) polymerase
Zinc finger, Zinc ‘POU’ domain signature
Kringle domain
Actinin-type actin-binding domain
Nuclear hormones
Eukaryotic putative transcription factor
‘four-disulfide core’ domain signature
family
or DNA
Recognition
Transcription factor
‘Cold-shock’ DNA-binding signature
family
or RNA
DNA
Bacterial activator proteins, Leucine zipper
Poly(ADP-ribose) polymerase
Zinc
‘POU’ domain signature
Kringle domain
Actinin-type actin-binding domain
Nuclear hormones receptors DNA-binding region signature
Eryf1-type zinc finger domain
Poly(ADP-ribose) polymerase zinc finger domain
Leucine zipper pattern
Fox/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
‘Cold-shock’ DNA-binding domain signature
CTF/NF-1 signature
ETS-domain signatures
SRF-type transcription factors DNA-binding and dimerization domain
Transcription factor TFHD repeat signature
eIF-4A family ATP-dependent helicases signatures
Eukaryotic putative RNA-binding region RNP-1 signature
Bacterial activator proteins, aroC family signature
Bacterial activator proteins, crp family signature
Bacterial activator proteins, gmr family signature
Bacterial activator proteins, lacI family signature
Bacterial activator proteins, lxrS family signature
Bacterial histone-like DNA-binding proteins signature
Histone H2A signature
Histone H2B signature
Histone H3 signature
Histone H4 signature
HMG1/2 signature
HMG-1 and HMG-Y DNA-binding domain (A + T-hook)
HMG14 and HMG17 signature
Protamine P1 signature
Ribosomal protein L5 signature
Ribosomal protein L11 signature
Ribosomal protein L14 signature
Ribosomal protein L23 signature
Ribosomal protein L39/L46 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 S15 signature
Ribosomal protein S17 signature
Ribosomal protein S18 signature
Ribosomal protein S19 signature
DNA mismatch repair proteins mutL / hexB / PMS1 signature

Enzymes

Oxidoreductases
Zinc-containing alcohol dehydrogenases signature
Iron-containing alcohol dehydrogenases signature
Insect-type alcohol dehydrogenase / ribitol dehydrogenase family signature
Aldo/keto reductase family signatures
L-lactate dehydrogenase active site
Glycerate-type 2-hydroxyacid dehydrogenases signature
Hydroxymethylglutaryl-coenzyme A reductases signatures
3-hydroxyacyl-CoA dehydrogenase signature
Malate dehydrogenase active site signature
Malic enzymes signature
Glucose-6-phosphate dehydrogenase active site
Bacterial quinoprotein dehydrogenases signatures
Aldehyde dehydrogenases active site
Glyceraldehyde 3-phosphate dehydrogenase active site
Acyl-CoA dehydrogenases signatures
Glutamate dehydrogenases active site
Dihydrofolate reductase signature
Pyridine nucleotide-disulphide oxidoreductases active site
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
Lipoxygenases, putative iron-binding region signature
Extradiol ring-cleavage dioxygenases signature
Biotin-dependent aromatic amino acid hydroxylases signature
Copper type II, ascorbate-dependent monooxygenases signatures
Cytochrome P450 cytisine heme-iron ligand signature
Copper/Zinc superoxide dismutase signature
Manganese and iron superoxide dismutases signature
Ribonucleotide reductase large subunit signature
Ribonucleotide reductase small subunit signature
Nitrogenases component 1 alpha and beta subunits signature

Transferrases
Thymidylate synthase active site
Methylnlated-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
Phosphobisphosphatidic acid formyltransferase active site
Aspartase and ornithine carbamoyltransferases signature
Thiolas signatures
Chloromethylphenyl acetiltransferase active site
CysE / lacA / nodL acetiltransferases signature
Phosphorylase pyridoxal-phosphate attachment site
UDP-glucoronosyl and UDP-glucosyl transfersases signature
Purine/pyrimidine phosphoribosyl transfersases signature
S-Adenosylmethionine synthetase signatures
EPSP synthase active site
Aspartate aminotransferases pyridoxal-phosphate attachment site
Hexokinases signature
Galactokinase signature
Hormones and active peptides

Adipokinetic hormone family signature
Bombesin-like peptides family signature
Calcitonin / CGRP / IAPP family signature
Chromogranins / secretogranins signatures
Gastrin / cholecystokinin family signature
Glucagon / GIP / secretin / VIP family signature
Glycoprotein hormones beta chain signature
Insulin family signature
Natriuretic peptides signature
Neurohypophysial hormones signature
Pancreatic hormone family signature
Parathyroid hormone family signature
Somatotropin, prolactin and related hormones signatures

Cytokines and growth factors

Int-1 family signature
HBGF/FGF family signature
Nerve growth factor family signature
Platelet-derived growth factor (PDGF) family signature
TGF-beta family signature
TNF family signature
Interferon alpha and beta family signature
Interleukin-1 signature
Interleukin-2 signature
Interleukin-6 / G-CSF / MGF signature
Interleukin-7 signature

Inhibitors

Pancreatic trypsin inhibitor (Kunitz) family signature
Bowman-Birk serine protease inhibitors family signature
Kazal serine protease inhibitors family signature
Soysbean trypsin inhibitor (Kunitz) protease inhibitors family signature
Serpins signature
Potato inhibitor 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
Disintegrins signature

Others

Pentraxin family signature
Immunoglobulins and major histocompatibility complex proteins signature
Prion protein signature
Cyclin signature
Proliferating cell nuclear antigen signature
Arrestins signature
Chaperonins signature
Heat shock hsp70 proteins family signatures
Heat shock hsp90 proteins family signature
Ubiquitin signature
SRP54-type proteins GTP-binding domain signature
GTP-binding elongation factors signature
Eukaryotic initiation factor 4D hypusine signature
S-100/CaBP type calcium binding protein signature
Hemolysin-type putative calcium-binding region signature
Small, acidic-soluble spore proteins, alpha/beta type, signature
Caseins alpha/beta signature
Legume lectins signatures
Vertebrate galactoside-binding lectin signature
Lysozyme-associated membrane glycoproteins signatures
Glycophorin A signature
Seminal vesicle protein I repeats signature
HCP repeats signature
Bacterial ice-nucleation proteins octamer repeat
Cell cycle proteins ftsW / rodA / spoVE signature
Staphylococcal coagulase repeat signature
11-S plant seed storage proteins signature
Dehydrins signature
Small hydrophilic plant seed proteins signature
Thaumatin family signature