DOMO: a new database of aligned protein domains

Jérôme Gracy, Patrick Argos

DOMO: a new database of aligned protein sequence domains.

Domains are autonomously folding units which are combined into modular proteins[1]. At a sequence level, accurately delineating the boundaries of homologous protein domains is essential for multiple sequence alignment. Tertiary structural data that could guide visual determination of such domain boundaries are not available for most proteins. Consequently, although many motif[2], block[3], and full-sequence-alignment[4] databases exist, as yet there are only two domain-alignment databases that have been constructed by a fully automated process utilizing only sequence information[5,6].

Here, we describe DOMO, a new database containing 8877 multiple sequence alignments, including 99 058 protein domains as well as repeating-sequence regions extracted from 83 054 non-redundant amino acid sequences from the SWISS-PROT[7] and PIR[8] databases. The domain boundaries and alignments were generated by a fully automated analysis process that involves the detection and clustering of amino acid sequence similarities and, subsequently, delineation of the domain boundaries and multiple sequence alignment of related protein segments[9,10]. The domain boundaries were not inferred from three-dimensional data. Instead, the relative positions of homologous segment pairs within the same protein (for repeats) or within homologous proteins with regard to each protein's N- or C-terminus were used to define the domain boundaries. The completeness and accuracy of the protein classifications, the correctness of the domain boundaries, and the quality of the multiple sequence alignments are greatly improved in DOMO, in comparison to other databases[9,10].

The database format

Each entry in the database corresponds to one family of homologous domains. Fields provide information about related proteins, functions of members of each family, decomposition of each protein into constituent domains, multiple sequence alignment, conserved residues and phylogenetic tree.

The multiple alignment shown in Fig. 1 includes 16 apple domains from four homologous serine proteases (four apple domains are present in each protein). It should be noted that the aligned fragments were delineated automatically and spliced from the complete sequences before multiple alignment.

Accessing DOMO through the World Wide Web

DOMO can be accessed through the sequence retrieval system SRS[11] at http://www.infobiogen.fr/services/domo/ which provides a form-based query manager allowing retrieval of famillial domain alignments by identifiers, sequence accession numbers, or keywords. Furthermore, the query results can be linked to other sequence databases to collect additional information on the relevant proteins. Domain families can also be translated to the standard MSF or FASTA formats so that they can be processed by other sequence-analysis tools.

If the three-dimensional structure of a domain is known, DOMO provides direct links to the corresponding atomic coordinates, RASMOL display[12], and complementary structural databases. If the domain structure is not known, the entry points to a composite prediction obtained by different techniques such as PREDATOR[13], SIMPA96[14], DSC[15], MAP123D[16], COILS[17], or

TMAP[18]. Moreover, a fast search for homologous domains can be performed using BLAST2[19], as can multiple sequence alignment using Clustal W[20]. The SRS steps in such an analysis are described below.

(1) From the SRS 'Select Libraries' page, select SWISSPROT or any other protein sequence database (but not DOMO).
(2) From the 'Query Form' page, formulate a sequence query.
(3) From the 'Query Result' page, launch BLAST2.
(4) From the 'BLAST2 options' page, select DOMO as the searched databank (and edit the sequence in query box if desired).
(5) From the 'BLAST2 Query Result' page, push the 'selected' button, select some hits, and launch Clustal W.
(6) From the 'Clustal W options' page, push the 'Align the selected hits with the query' button if desired.
(7) If the protein has multiple domains, return to 'BLAST2 Query Result' page, and reiterate steps (5) and (6) with hits from not already aligned DOMO families.

Concluding remarks

The DOMO sequence-analysis environment (domain database, query manager, homology search and multiple-sequence-alignment tools) provides a simple tool for determining domain arrangements, evolutionary relationships, and key amino acid residues in a query protein sequence.

References

1 Rossmann, M. G. and Argos, P. (1981) *Annu. Rev. Biochem.* 50, 497-532
2 Bairoch, A., Bucher, P. and Hofmann, K. (1996) Nucleic Acid Res. 24, 189-196
3 Henikoff, J. G. and Henikoff, S. (1996) *Methods Enzymol.* 266, 88-105
4 Gonnet, G., Cohen, M. A and Benner, S. A. (1992) *Science* 256, 1443-1445
5 Sonnhammer, E. L. and Kahn, D. (1994) *Protein Sci.* 3, 482-492
6 Sonnhammer, E. L., Eddy, S. R. and Durbin, R. (1997) *Proteins* 28, 405-420
7 Bairoch, A. and Apweiler, R. (1996) *Nucleic Acid Res.* 24, 21-25
8 George, D. G. et al. (1996) *Nucleic Acid. Res.* 24, 17-20
9 Gracy, J. and Argos, P. (1998) *Bioinformatics* 14, 163-173
10 Gracy, J. and Argos, P. (1998) *Bioinformatics* 14, 174-187
11 Etzold, T., Ulyanov, A. and Argos, P. (1996) *Methods Enzymol.* 266, 114-128
12 Sayle, R. A. and Milner-White, E. J. (1995) *Trends Biochem. Sci.* 20, 374
13 Frishman, D. and Argos, P. (1996) *Protein Eng.* 9, 133-142
14 Levin, J. M. (1997) *Protein Eng.* 10, 771-776
15 King, R. D. and Sternberg, M. J. E. (1996) *Protein Science* 5, 2298-2310
16 Gracy, J., Chiche, L. and Sallantin, J. (1993) *Biochimie* 75, 353-361
17 Lupas, A. (1996) *Trends Biochem. Sci.* 21, 375-382
18 Persson, B. and Argos, P. (1994) *J. Mol. Biol.* 237, 182-192
19 Altschul, S. F et al. (1997) *Nucleic Acids Res.* 25, 3389-3402
20 Higgins, D. G., Thompson, J. D. and Gibson, T. J. (1996) *Methods Enzymol.* 266, 383-402

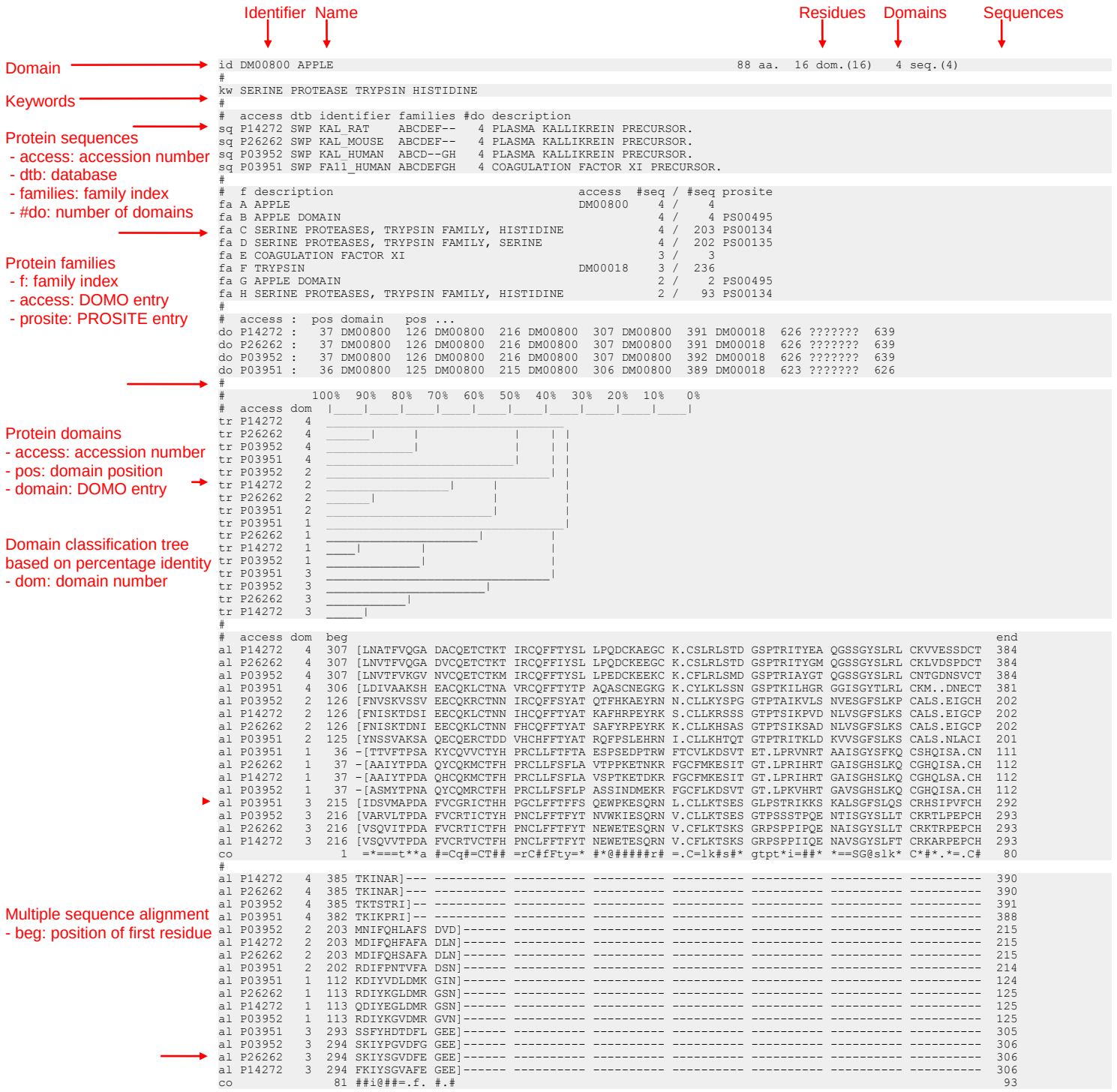
Jérôme GRACY

Infobiogen, 7 rue Guy Môquet, 94801 Villejuif Cedex, France.

Patrick ARGOS
EMBL, Meyerhof Strasse 1, Postfach 10.2209, D-69012 Heidelberg, Germany.

Figure 1

Example of a DOMO database entry : the apple domain.
The consensus sequence generated shows residues as lowercase characters if the conservation level is above 85% and below 100%, and as uppercase characters if the position is absolutely conserved. Physico-chemical properties are also shown (e.g. positions that are mainly hydrophobic or hydrophilic residues are indicated by the signs = and #, respectively).




```
id DM00800 APPLE                                                    88 aa.  16 dom.(16)   4 seq.(4)
#
kw SERINE PROTEASE TRYPSIN HISTIDINE
#
#  access dtb identifier families #do description
sq P14272 SWP KAL_RAT    ABCDEF--   4 PLASMA KALLIKREIN PRECURSOR.
sq P26262 SWP KAL_MOUSE  ABCDEF--   4 PLASMA KALLIKREIN PRECURSOR.
sq P03952 SWP KAL_HUMAN  ABCD--GH   4 PLASMA KALLIKREIN PRECURSOR.
sq P03951 SWP FA11_HUMAN ABCDEFGH   4 COAGULATION FACTOR XI PRECURSOR.
#
#  f description                                  access  #seq / #seq prosite
fa A APPLE                                        DM00800    4 /    4
fa B APPLE DOMAIN                                            4 /    4 PS00495
fa C SERINE PROTEASES, TRYPSIN FAMILY, HISTIDINE             4 /  203 PS00134
fa D SERINE PROTEASES, TRYPSIN FAMILY, SERINE                4 /  202 PS00135
fa E COAGULATION FACTOR XI                                   3 /    3
fa F TRYPSIN                                      DM00018    3 /  236
fa G APPLE DOMAIN                                            2 /    2 PS00495
fa H SERINE PROTEASES, TRYPSIN FAMILY, HISTIDINE             2 /   93 PS00134
#
#  access :  pos domain   pos ...
do P14272 :   37 DM00800  126 DM00800  216 DM00800  307 DM00800  391 DM00018  626 ???????  639
do P26262 :   37 DM00800  126 DM00800  216 DM00800  307 DM00800  391 DM00018  626 ???????  639
do P03952 :   37 DM00800  126 DM00800  216 DM00800  307 DM00800  392 DM00018  626 ???????  639
do P03951 :   36 DM00800  125 DM00800  215 DM00800  306 DM00800  389 DM00018  623 ???????  626
#
#            100%  90%  80%  70%  60%  50%  40%  30%  20%  10%   0%
#  access dom  |____|____|____|____|____|____|____|____|____|____|
tr P14272   4  _____________________________________
tr P26262   4  ______|     |               |     | |
tr P03952   4  _____________|               |     | |
tr P03951   4  ____________________________|     | |
tr P03952   2  ___________________________________| |
tr P14272   2  ___________________|     |          |
tr P26262   2  ______|               |          |
tr P03951   2  ____________________|          |
tr P03951   1  ______________________________|
tr P26262   1  ___________________________|          |
tr P14272   1  ____|         |                 |
tr P03952   1  _____________|                 |
tr P03951   3  ______________________________|
tr P03952   3  ___________________________|
tr P26262   3  ___________|
tr P14272   3  _____|
#
#  access dom  beg                                                                                end
al P14272   4  307 [LNATFVQGA DACQETCTKT IRCQFFTYSL LPQDCKAEGC K.CSLRLSTD GSPTRITYEA QGSSGYSLRL CKVVESSDCT  384
al P26262   4  307 [LNVTFVQGA DVCQETCTKT IRCQFFIYSL LPQDCKEEGC K.CSLRLSTD GSPTRITYGM QGSSGYSLRL CKLVDSPDCT  384
al P03952   4  307 [LNVTFVKGV NVCQETCTKM IRCQFFTYSL LPEDCKEEKC K.CFLRLSMD GSPTRIAYGT QGSSGYSLRL CNTGDNSVCT  384
al P03951   4  306 [LDIVAAKSH EACQKLCTNA VRCQFFTYTP AQASCNEGKG K.CYLKLSSN GSPTKILHGR GGISGYTLRL CKM..DNECT  381
al P03952   2  126 [FNVSKVSSV EECQKRCTNN IRCQFFSYAT QTFHKAEYRN N.CLLKYSPG GTPTAIKVLS NVESGFSLKP CALS.EIGCH  202
al P14272   2  126 [FNISKTDSI EECQKLCTNN IHCQFFTYAT KAFHRPEYRK S.CLLKRSSS GTPTSIKPVD NLVSGFSLKS CALS.EIGCP  202
al P26262   2  126 [FNISKTDNI EECQKLCTNN FHCQFFTYAT SAFYRPEYRK K.CLLKHSAS GTPTSIKSAD NLVSGFSLKS CALS.EIGCP  202
al P03951   2  125 [YNSSVAKSA QECQERCTDD VHCHFFTYAT RQFPSLEHRN I.CLLKHTQT GTPTRITKLD KVVSGFSLKS CALS.NLACI  201
al P03951   1   36 -[TTVFTPSA KYCQVVCTYH PRCLLFTFTA ESPSEDPTRW FTCVLKDSVT ET.LPRVNRT AAISGYSFKQ CSHQISA.CN  111
al P26262   1   37 -[AAIYTPDA QYCQKMCTFH PRCLLFSFLA VTPPKETNKR FGCFMKESIT GT.LPRIHRT GAISGHSLKQ CGHQISA.CH  112
al P14272   1   37 -[AAIYTPDA QHCQKMCTFH PRCLLFSFLA VSPTKETDKR FGCFMKESIT GT.LPRIHRT GAISGHSLKQ CGHQLSA.CH  112
al P03952   1   37 -[ASMYTPNA QYCQMRCTFH PRCLLFSFLP ASSINDMEKR FGCFLKDSVT GT.LPKVHRT GAVSGHSLKQ CGHQISA.CH  112
al P03951   3  215 [IDSVMAPDA FVCGRICTHH PGCLFFTFFS QEWPKESQRN L.CLLKTSES GLPSTRIKKS KALSGFSLQS CRHSIPVFCH  292
al P03952   3  216 [VARVLTPDA FVCRTICTYH PNCLFFTFYT NVWKIESQRN V.CLLKTSES GTPSSSTPQE NTISGYSLLT CKRTLPEPCH  293
al P26262   3  216 [VSQVITPDA FVCRTICTFH PNCLFFTFYT NEWETESQRN V.CFLKTSKS GRPSPPIPQE NAISGYSLLT CRKTRPEPCH  293
al P14272   3  216 [VSQVVTPDA FVCRTVCTFH PNCLFFTFYT NEWETESQRN V.CFLKTSKS GRPSPPIIQE NAVSGYSLFT CRKARPEPCH  293
co              1  =*===t**a #=Cq#=CT## =rC#fFty=* #*@#####r# =.C=lk#s#* gtpt*i=##* *==SG@slk* C*#*.*=.C#   80
#
al P14272   4  385 TKINAR]--- ---------- ---------- ---------- ---------- ---------- ---------- ----------  390
al P26262   4  385 TKINAR]--- ---------- ---------- ---------- ---------- ---------- ---------- ----------  390
al P03952   4  385 TKTSTRI]-- ---------- ---------- ---------- ---------- ---------- ---------- ----------  391
al P03951   4  382 TKIKPRI]-- ---------- ---------- ---------- ---------- ---------- ---------- ----------  388
al P03952   2  203 MNIFQHLAFS DVD]------ ---------- ---------- ---------- ---------- ---------- ----------  215
al P14272   2  203 MDIFQHFAFA DLN]------ ---------- ---------- ---------- ---------- ---------- ----------  215
al P26262   2  203 MDIFQHSAFA DLN]------ ---------- ---------- ---------- ---------- ---------- ----------  215
al P03951   2  202 RDIFPNTVFA DSN]------ ---------- ---------- ---------- ---------- ---------- ----------  214
al P03951   1  112 KDIYVDLDMK GIN]------ ---------- ---------- ---------- ---------- ---------- ----------  124
al P26262   1  113 RDIYKGLDMR GSN]------ ---------- ---------- ---------- ---------- ---------- ----------  125
al P14272   1  113 QDIYEGLDMR GSN]------ ---------- ---------- ---------- ---------- ---------- ----------  125
al P03952   1  113 RDIYKGVDMR GVN]------ ---------- ---------- ---------- ---------- ---------- ----------  125
al P03951   3  293 SSFYHDTDFL GEE]------ ---------- ---------- ---------- ---------- ---------- ----------  305
al P03952   3  294 SKIYPGVDFG GEE]------ ---------- ---------- ---------- ---------- ---------- ----------  306
al P26262   3  294 SKIYSGVDFE GEE]------ ---------- ---------- ---------- ---------- ---------- ----------  306
al P14272   3  294 FKIYSGVAFE GEE]------ ---------- ---------- ---------- ---------- ---------- ----------  306
co              81 ##i@##=.f. #.#                                                                   93
```