200–202 *Nucleic Acids Research, 2002, Vol. 30, No. 1*

ASPD (Artificially Selected Proteins/Peptides Database): a database of proteins and peptides evolved *in vitro*

Vadim P. Valuev*, Dmitry A. Afonnikov, Mikhail P. Ponomarenko, Luciano Milanesi[1] and Nikolay A. Kolchanov

Institute of Cytology and Genetics, Lavrentieva ave. 10, Novosibirsk 630090, Russia and [1]Istituto Tecnologie Biomediche Avanzate, Via Fratelli Cervi 93, 20090 Segrate, Italy

Received August 22, 2001; Revised and Accepted October 2, 2001

ABSTRACT

ASPD is a new curated database that incorporates data on full-length proteins, protein domains and peptides that were obtained through *in vitro* directed evolution processes (mainly by means of phage display). At present, the ASPD database contains data on 195 selection experiments, which were described in 112 original papers. For each experiment, the following information is given: (i) description of the target for binding, (ii) description of the protein or peptide which serves as the template for library construction and description of the native protein which binds the target, (iii) links to the major proteomic databases (SWISS-PROT, PDB, PROSITE and ENZYME), (iv) keywords referring to the biological significance of the experiment, (v) aligned sequences of proteins or peptides retrieved through *in vitro* evolution and relevant native or constructed sequences, (vi) the number of rounds of selection/amplification and (vii) the number of occurrences of clones with each sequence. The literature data include a full reference, a link to the MEDLINE database and the name of the corresponding author with his email address. ASPD has a user-friendly interface which allows for simple queries using the names of proteins and ligands, as well as keywords describing the biological role of the interaction studied, and also for queries based on authors' names. It is also possible to access the database by means of the SRS system, allowing complex queries. There is a BLAST search tool against the ASPD for looking directly for homologous sequences. Research tools of the ASPD allow the analysis of pairwise correlations in the sequences of proteins and peptides selected against one target. The URL for the ASPD database is http://www.sgi.sscc.ru/mgs/gnw/aspd/.

INTRODUCTION

The early 1990s saw the start of experiments applying the technique of *in vitro* evolution of nucleic acids and proteins. This process involves sieving large pools of molecules (up to 10^{13} individual members) through several consecutive rounds of selection and amplification to retrieve those molecules that show the desired property (1,2). It has a very wide range of applications including: studying protein folding and thermodynamic stability, engineering proteins with new or improved enzymatic activities (including catalytic antibodies), mapping epitopes and binding sites, finding substrates for enzymes, raising highly specific antibodies and finding protein mimics for non-protein molecules. That such a wide field is covered by these experiments illustrates the great promise and importance of data obtained by *in vitro* protein evolution, and simultaneously makes it very difficult to establish a uniform way of describing all the experiments.

ASPD is a new curated and annotated database which aims to arrange the data being obtained by means of *in vitro* protein evolution into a structured and easily searchable array, and to integrate it into the network of molecular biological resources. This could simplify the work of researchers carrying out new work in this field, as well as aid in protein annotation and structure–function analysis.

FORMAT OF THE ASPD DATABASE

The database consists of two tables (sets of entries) called ASPD_ALIGN and ASPD_REF. Each entry in ASPD_ALIGN contains one annotated alignment of selected *in vitro* proteins or peptides, which usually corresponds to one selection experiment. ASPD_REF contains literature references. Each entry of ASPD_ALIGN has a link to the corresponding entry in ASPD_REF. Each field in the two tables starts with its name, and if it is continued on the next line, this line starts with a blank space. Below we give a detailed description of the fields in the order of their appearance.

Fields of the ASPD_ALIGN table

Identifier. This is the only unique identifier of an entry. It follows the pattern PH1XXNNN, where X is a letter, and N is a number.

Lit_reference. An identifier of the corresponding entry in the ASPD_REF table (literature reference). It appears as a hyperlink.

*To whom correspondence should be addressed. Tel: +7 383 233 2971; Fax: +7 383 233 1278; Email: valuev@bionet.nsc.ru

Target. An informal description of the target for binding. It usually comprises the name of the target, its type (protein, other organic or inorganic molecule, DNA, cell), organism name (if applicable), a brief description of its biological function and some specific conditions of the experiment, if necessary. Some examples of the 'target' field include:

1. Toxic shock syndrome toxin 1 (TSST-1), staphylococcal enterotoxin (*Staphylococcus aureus*).
2. Glutathione-*S*-transferase (GST).
3. Tumor-associated antigen polymorphic epithelial mucin (MUC1), transmembrane molecule present at the apical cell surface of normal secretory epithelial tissues of breast, ovary, colon, lung, pancreas and acts as a protective barrier.
4. Mouse kidney.
5. 8-Oxoguanine.

Link. This is a link to external databases: SWISS-PROT (3), PDB (4), PROSITE (5) and ENZYME (6). It has the following form: 'DB_Name DB_Ref', where DB_Name is the name of the database (one of those listed above) and DB_Ref is the identifier of the entry in this database. The field 'Link' is multiple and follows the field 'Target' if the corresponding entry in the external database pertains to the target for binding, and the field 'Template' if it pertains to the protein or peptide subject to randomization and selection.

Template. Description of the protein or peptide which was partly or fully randomized, or that of the native protein for which mimics are sought. This description is made in the same way as in the 'Target' field. Examples of the field 'Template':

1. Peptides mimicking lipopolysaccharide (LPS), which is the virulence determinant of *Brucella*.
2. Cyclic random 11-amino acid peptides mimicking pTyr-containing motifs.
3. Plantago major (common plantain) sucrose carrier protein, N-terminus.

Keywords. Keywords refer to the nature of interaction and interacting partners, as well as to its biological significance. They cover from broad (such as 'disease' or 'antibody') to narrow (e.g. 'Hodgkin's disease') subjects. Examples of the most frequent keywords are 'epitope mapping', 'signal transduction', 'cell adhesion', 'differentiation' and 'membrane'.

Comment. The comment usually refers to some peculiarities of the experiment, for example that different libraries were used or that the conditions for selection were not exactly the same for all sequences. This field is not searchable.

Consensus. The author-derived consensus of the randomized stretch. If given, it is in PROSITE format. This field is not searchable.

Number_of_sequences. The number of different amino acid sequences reported in the experiment described.

Alignment. This field contains information about the amino acid sequences of proteins and peptides obtained in the given experiment, along with some details of the experiment itself. The full-length sequences of proteins and peptides retrieved *in vitro* are shown (corresponding to the entire molecule exposed on the phage surface), even if they were partly randomized and only the amino acids in positions subject to randomization are reported in the original paper. All the sequences are numbered and they are aligned. At the end of the line with a sequence, the number of clones with this sequence is given. The amino acids in the randomized positions are shown in upper case; those in the non-randomized positions are in lower case. Below the sequences retrieved *in vitro* we give the relevant sequences of native or constructed proteins. These sequences are marked with latin letters and are in upper case. Finally, in the first line of the 'Alignment' field, the number of selection/amplification cycles is shown.

Fields of the ASPD_REF table

The table ASPD_REF contains the following fields: Identifier, Authors, Title, Journal, Volume, Year, Pages, MEDLINE and Corresponding_Author.

ACCESS TO THE ASPD

The ASPD web site is located at http://www.sgi.sscc.ru/mgs/gnw/aspd/. From this web page, the user can perform simple queries to the database. Two kinds of simple query are possible: the first uses the description of the experiment (the fields 'Target', 'Template' and 'Keywords' are searched); whilst the second uses author name. A list of short descriptions of the entries found results from these searches. These descriptions include information on binding targets (field 'Target'), the protein or peptide under study (field 'Template') and the list of authors of the corresponding paper. From these descriptions the user can proceed to the full entries by means of hyperlinks. To implement more complex and exact queries one can access the ASPD via SRS6 (Sequence Retrieval System) (7). There are links on the ASPD web page to the SRS6 pages, from which one can proceed to the query forms (by clicking the 'Search' button) using data on experiment (ASPD_ALIGN) and literature references (ASPD_REF). Within these forms the user can build complex queries by selecting the database fields and expressions to search for in these fields and combining them with logical operators. The user can define the output representation (which fields to show) from these query forms as well. This entire operation of query building requires just the use of the boxes and selection from the lists present on the query page. The simplest example of an SRS search is when one wants to search for some expression in an exact field (for example, for 'streptavidin' in the field 'Target'). (See more details of SRS use in the Supplementary Material.) Finally, to handle protein sequences directly, we provide the possibility of BLAST searching (8) against ASPD. Following the link on the ASPD web page (http://www.sgi.sscc.ru/mgs/systems/fastprot/aspd_blast.html), one can get to a BLAST search input form from where it is possible to submit the sequence of interest, set the appropriate search parameters (or leave the default values) and discover whether ASPD contains any sequences homologous to the one submitted. From the resulting BLAST page there are the links back to the original ASPD entries.

DATA SUBMISSION

People carrying out *in vitro* selection experiments are welcome to submit their data, unpublished as well as published, by filling in the data submission form available from the ASPD web page and sending it to the ASPD administrator at valuev@bionet.nsc.ru. All comments (sent to the same address) are also welcome.

DATABASE STATISTICS

At present, ASPD contains descriptions of 195 experiments based on data from 112 references, which correspond to 4345 sequences of proteins and peptides selected *in vitro*. In total there are 395 links to SWISS-PROT, 251 links to PDB, 140 links to PROSITE and 32 links to ENZYME.

RESEARCH BASED ON DATA STORED IN ASPD

We have previously developed the CRASP system (9) (http://www.sgi.sscc.ru/mgs/gnw/crasp/), a tool for analyzing correlated substitutions in pairs of positions in related proteins. If the value of some property of the amino acid in a given position in a number of related proteins correlates with the value of the same property in another position in the same proteins, that may mean that these two amino acids interact in some way. The simplest explanation of such an interaction is for the two amino acids to be adjacent in the protein's three-dimensional structure. We applied CRASP to calculate such pairwise correlations for all aligned sets of proteins and peptides (each aligned set corresponds to an ASPD entry) in terms of four amino acid properties: hydrophobicity (10), volume (11), isoelectric point value (12) and polarity (13). The data on these correlations are available both in text and graphical formats by following the link 'Correlation analysis' from each ASPD entry. We have also calculated the amino acid compositions of *in vitro* evolved proteins and peptides (taking into account only amino acids in positions that were randomized) and amino acid similarity matrices following the procedure described previously (14). These data are available from the ASPD web page. More details about the research are given in the Supplementary Material.

FUTURE DEVELOPMENTS

We are continuing to expand the ASPD database. It is updated on a regular basis, approximately once every 2 months. In the next year we are going to double its size so that it will embrace the great majority of published data.

Some improvements in its format are also imminent. Short-term goals include: formalizing the descriptions of targets for binding and templates for randomization to make the data more structured, clear and searchable; and enabling the visualization of three-dimensional structures of those proteins and protein–ligand complexes studied by *in vitro* evolution. The medium-term goal is to implement a facility to search the SWISS-PROT and TrEMBL (3) databases for proteins and domains homologous to those stored in ASPD. Finally, the long-term goal is to include more details relating to the experiments themselves [description of the randomized library, type of phage and phage protein used (in the case of phage display), strategies used for elution, etc.].

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We are grateful to Dr Jonathan M. Blackburn and Wayne Patrick (both University of Cambridge, UK) for comments on the manuscript and assistance in its preparation in English. We would like to thank also Irena V. Lohova and Dr Vladimir A. Ivanisenko for help in work and valuable discussions. This work was supported by grant nos 00-04-49229 (and its subgrant no. 01-04 06240) and 99-04-49879 from the Russian Foundation for Basic Research and grant YSF 00-177 from INTAS.

REFERENCES

1. Roberts,R.W. and Ja,W.W. (1999) *In vitro* selection of nucleic acids and proteins: what are we learning? *Curr. Opin. Struct. Biol.*, **9**, 521–529.
2. Smith,G.P. and Scott,J.K. (1993) Libraries of peptides and proteins displayed on filamentous phage. *Methods Enzymol.*, **217**, 228–257.
3. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
4. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 245–248.
5. Hofmann,K., Bucher,P., Falquet,L. and Bairoch,A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.*, **27**, 215–219. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 235–238.
6. Bairoch,A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.*, **28**, 304–305.
7. Etzold,T. and Argos,P. (1993) SRS – an indexing and retrieval tool for flat file data libraries. *Comput. Appl. Biosci.*, **9**, 49–57.
8. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search program. *Nucleic Acids Res.*, **25**, 3389–3402.
9. Afonnikov,D.A., Oshchepkov,D.Yu. and Kolchanov,N.A. (2001) Detection of conserved physico-chemical characteristics of proteins by analysing clusters of positions with co-ordinated substitutions. *Bioinformatics*, in press.
10. Eisenberg,D., Schwarz,E., Komaromy,M. and Wall,R. (1984) Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.*, **179**, 125–142.
11. Chothia,C. (1984) Principles that determine the structure of proteins. *Annu. Rev. Biochem.*, **53**, 537–572.
12. Zimmerman,J.M., Eliezer,N. and Simha,R. (1968) The characterization of amino acid sequences in proteins by statistical methods. *J. Theor. Biol.*, **21**, 170–201.
13. Ponnuswamy,P.K., Prabhakaran,M. and Manavalan,P. (1980) Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins. *Biochim. Biophys. Acta*, **623**, 301–316.
14. Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.