

DB-PABP: a database of polyanion-binding proteins

Jianwen Fang^{1,2,*}, Yinghua Dong¹, Nazila Salamat-Miller³ and C. Russell Middaugh⁴

¹Bioinformatics Core Facility, ²Information and Telecommunication Technology Center, University of Kansas, Lawrence, KS 66047, ³Shire Human Genetic Therapeutics, Cambridge, MA 02139 and ⁴Department of Pharmaceutical Chemistry, University of Kansas, Lawrence, KS 66047, USA

Received August 10, 2007; Revised September 12, 2007; Accepted September 17, 2007

ABSTRACT

The interactions between polyanions (PAs) and polyanion-binding proteins (PABPs) have been found to play significant roles in many essential biological processes including intracellular organization, transport and protein folding. Furthermore, many neurodegenerative disease-related proteins are PABPs. Thus, a better understanding of PA/PABP interactions may not only enhance our understandings of biological systems but also provide new clues to these deadly diseases. The literature in this field is widely scattered, suggesting the need for a comprehensive and searchable database of PABPs. The DB-PABP is a comprehensive, manually curated and searchable database of experimentally characterized PABPs. It is freely available and can be accessed online at http://pabp.bcf.ku.edu/DB_PABP/. The DB-PABP was implemented as a MySQL relational database. An interactive web interface was created using Java Server Pages (JSP). The search page of the database is organized into a main search form and a section for utilities. The main search form enables custom searches via four menus: protein names, polyanion names, the source species of the proteins and the methods used to discover the interactions. Available utilities include a commonality matrix, a function of listing PABPs by the number of interacting polyanions and a string search for author surnames. The DB-PABP is maintained at the University of Kansas. We encourage users to provide feedback and submit new data and references.

INTRODUCTION

Polyanion-binding proteins (PABPs) are a group of very diverse proteins distinguished by their physical interactions with polyanions (PAs). As the name indicates, polyanions are molecular entities bearing multiple negative charges. Most common polyanionic macromolecules

and macromolecular complexes which include proteoglycans, DNA, RNA, actin (microfilaments), tubulin (microtubules), polysialic acids, ribosomes, etc., are of extreme polyanionic nature and are widely dispersed throughout cells (1). Polyanions are involved in many essential biological processes such as: (a) regulatory functions; (b) generic information transfer; (c) protein folding and stabilization; and (d) transport. In addition, lower molecular weight polyanions such as nucleotides, phosphorylated inositols and polyphosphates bind to such proteins and play important regulatory roles. Recently it was discovered that many neurodegenerative disease-related proteins, such as those associated with Alzheimer's, Parkinson's and Prion diseases, are PABPs (2). Thus a better understanding of PA/PABP interactions may not only enhance our understandings of biological systems but also provide new clues to these deadly diseases and new targets for therapy (1–4).

Extensive efforts have been devoted to characterizing PABPs and their interactions with polyanions (5–8). The vast majority of these studies, however, have focused on only one or a few PA/PABP interactions under the assumption of a high degree of specificity, despite the finding that most PABPs can interact with many different polyanions (3,4). Although heparin is the prototypic polyanion (3), the designation on 'heparin-binding protein' is often a misnomer (5). These observations suggest a number of fundamental questions. Is there a network of PA/PABP interactions? If so, what are the global roles of such a network in living systems? We believe such questions can only be addressed with systematic studies based on a large collection of PA/PABP interactions.

Recently we used human and yeast protein arrays to identify a large number of PABPs interacting with one or more of five model polyanions: actin, tubulin, DNA, heparin and heparan sulfate (3,4). We also provided evidences for the existence of a network-like system for PABPs within cells and their potential roles as critical hubs in intracellular behavior. This network probably interlaces with protein–protein interaction networks, in which both proteins and polyanions act as interacting nodes (1,3,4). Other notable systematic studies include a large-scale identification of tubulin-binding proteins in

*To whom correspondence should be addressed. Tel: +1 785 864 3349; Fax: +1 785 864 5738; Email: jwfang@ku.edu

Arabidopsis (9) and heparin-binding proteins in human plasma (10) and *Escherichia coli* K-12 MG1655 cells (<http://eep.tamu.edu/heparome/>). These initial investigations have taken first steps toward achieving a better understanding of the nature of PA/PABP interactions within cells and provide a basis for future datamining studies (11). Current high throughput technologies, however, can only describe a portion of the PABPs found in an organism. For example, the human protein arrays used in the previous study only contained about 5000 proteins, a rather small fraction of estimated a hundred thousand proteins in human cells (3). Nevertheless, many PABPs and their interactions with polyanions have been documented and new instances are being described at a high rate in the primary literature. Thus a well-maintained, comprehensive database of PABPs including high throughput data as well as curated information from the literature is in need. Although there exist several databases of DNA-binding proteins (12–15) and heparin-binding proteins (<http://eep.tamu.edu/heparome/>), to the best of our knowledge, no comprehensive database of PABPs has been developed. Thus, we have built and are maintaining the DB-PABP to document publicly available, experimentally determined PABPs.

DATABASE CONSTRUCTION

Database and its web interface design

The DB-PABP was implemented as a MySQL relational database. An interactive web interface was created using Java Server Pages (JSP) and the Java Database Connectivity (JDBC) API was used to interface with the MySQL database. The schema of the database design is available on the website. We also used Perl scripts to retrieve information directly from the NCBI proteins database and PubMed. Currently our server runs on the Window 2003 Server operating system.

Populating the database

The information in the DB-PABP has been collected from original literature reports of experimentally verified PA/PABP interactions. PABPs identified with protein array technology were assigned with low, medium or high confidence levels, which correspond to a mean greater than a median signal of all proteins spots on the array plus one, two or three times the standard deviation, respectively (3,4). In an attempt to minimize the occurrence of false positives, we chose to include only PABPs with high confidence levels.

A simple but efficient data entry routine was implemented to facilitate populating the database. At each entry cycle the names of the polyanion and its PABP partner, the source species of the protein, the identification method employed and the original literature reference are entered. Various technologies have been employed to identify PABPs (16,17). To ensure data consistency and increase the efficiency and accuracy of searches, we used controlled vocabulary and ontology of interaction detection methods available in the Ontology Lookup Service (18). The DB-PABP uses protein annotations

from the NCBI protein database and literature information is retrieved from PubMed. In-house Perl scripts utilizing BioPerl modules (<http://www.bioperl.org/>) are used to retrieve protein and reference information from the NCBI protein database and PubMed based on NCBI accession numbers and PubMed IDs. The data entry routine is accessible to all registered users to allow them to contribute new references and data. To maintain data integrity, newly entered data are not available to the general public until each entry is double-checked against the original literature. A public/private flag is set to control the data accession prior to incorporating it into the public database. A private version of the search form, which can be used to search both public and private data sets, is available for the curators and registered users. With these data validation steps, we believe that the data contained in the DB-PABP are highly reliable.

Current status and updates

Up-to-date statistics of the DB-PABP are available in the website. As of 10 September 2007, the database has about 500 distinct PABPs involved in over 710 PA/PABP interactions. The information was extracted from more than 200 literature papers and only experimentally verified PABPs are considered in the database. We have been actively populating the database and plan to maintain at least weekly updates for the years to come. We encourage users to provide feedback and submit new data and references.

Currently most PABPs in the database are from human and yeast and are based upon their interactions with one or more of five representative polyanions (actin, tubulin, DNA, heparin and heparan sulfate). In the future we will extend the lists to other organisms and common polyanions. Since there already exist several transcription factors and DNA-binding proteins (12–15), at present we focus on newly identified DNA-binding proteins (3,4) and those with solved protein–DNA complex structures. In the near future, DB-PABP will include more DNA-binding proteins from other databases.

Main search form

The web-accessible search page is organized into a main search form and a section of utilities. The main search form enables custom searches via four menus: protein names, polyanion names, species and the methods used to discover the interactions. The polyanion menu supports ‘OR’ (one or more selected polyanions) and ‘AND’ (all selected polyanions) Boolean operations. The species and method menus only use ‘OR’ operation and by default are set to ‘any (all)’. Custom searches can be performed by selecting different combinations of these menus. The protein name menu can be refined by typing any part of a protein name of interest and then clicking the ‘check’ button. Multiple choices can be made in all four menus by holding the ‘Ctrl’ key.

The main search form generates tables containing specific information and hyperlinks (Figure 1). For example, typing ‘glycosylase’ and then clicking the ‘check’ button generates a short list of proteins whose

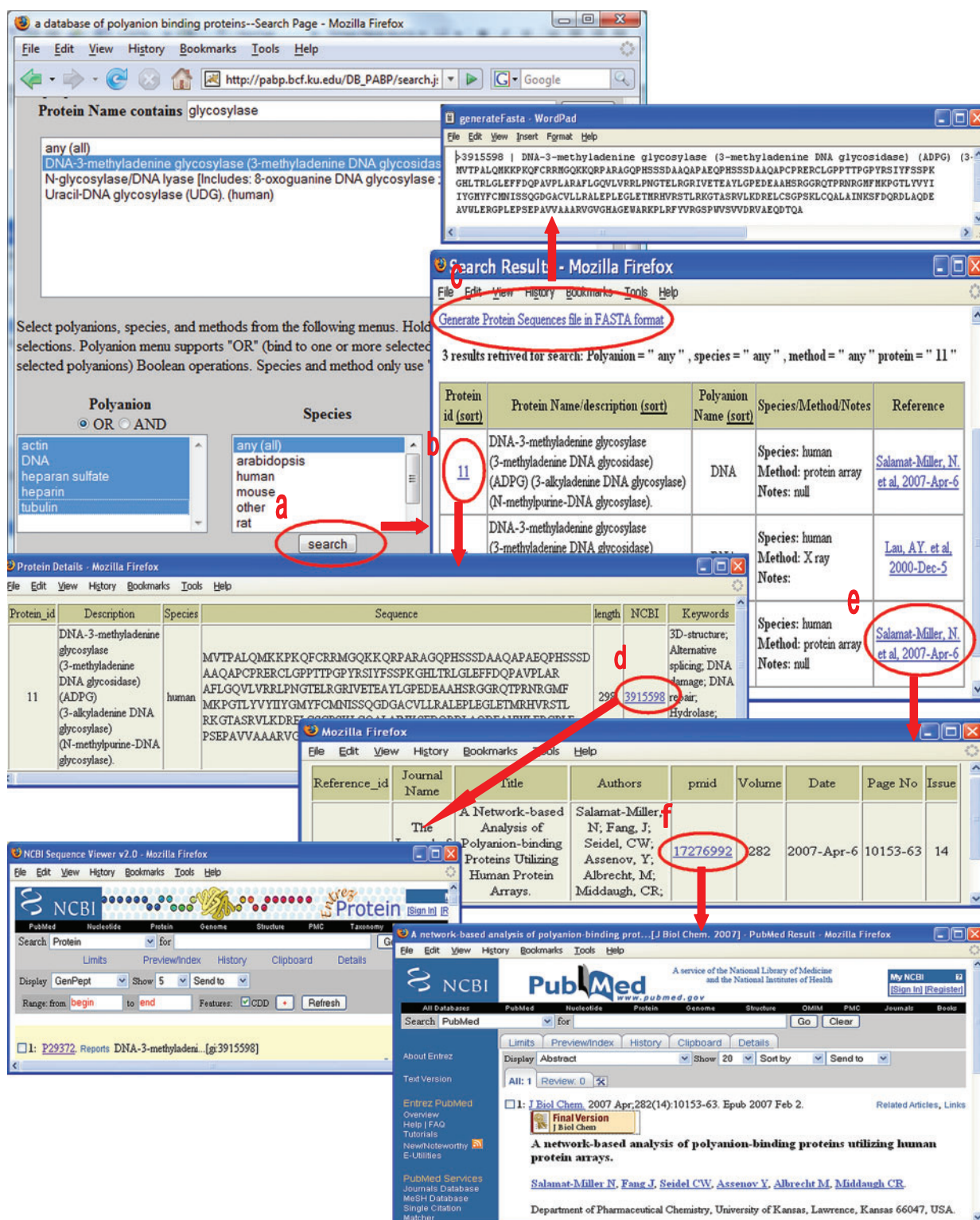


Figure 1. Examples of search outputs. (a) Main search form; (b) link to protein information; (c) generating a FASTA file; (d) link to NCBI protein database; (e) link to literature information; (f) link to PubMed.

names match 'glycosylase'. Selecting 'DNA-3-methyladenine glycosylase' from the protein menu and all five choices from the polyanion menu, keeping other menus with default choices, followed by clicking the 'search' button produces a table containing the following

information: (a) protein ID in the database (as an active hyperlink to protein information available in the database including a link to the NCBI protein database); (b) name and description of the proteins; (c) polyanion name; (d) a summary of the species, the identification

method used, and notes if any; and (e) a reference in the form of a hyperlink to the full citation information including the PubMed ID number (as an active hyperlink). Links to the NCBI protein database and PubMed are provided in the relevant output tables so users may access additional information in these public databases. A hyperlink on the top of the report returns a FASTA page listing all distinct PABPs in the report.

Utilities

In addition to the main search form, the DB-PABP provides a set of utilities which allow string searches for author surnames. A list of papers published by authors whose surnames match the input string will be returned as search results. The 'Commonality Matrix' function produces an $N \times N$ matrix, where N is the number of polyanions in the database (thus, currently N is 5). Each diagonal element shows the number of proteins known to interact with the polyanion whose row and column intersect on the diagonal. The off-diagonal elements display the number of proteins interacting with both of the polyanions whose row and column intersect at that off-diagonal element. Each cell in the commonality matrix is also a hyperlink that leads to a table providing information about all of the relevant PABPs.

Another useful utility, 'list proteins by number of hits', ranks all proteins by the number of their interacting polyanion partners. By default, the result table is sorted by the number of polyanions. It can also be resorted by protein ID, protein name/description and species.

AVAILABILITY AND REQUIREMENTS

The database is freely accessible at http://pabp.bcf.ku.edu/DB_PABP/. It has been tested and works with Mozilla Firefox 2 and Internet Explorer 5/6. Some features may not work well with other browsers (e.g. Mac Safari 2.0).

ACKNOWLEDGEMENTS

This work was supported in part by K-INBRE Bioinformatics Core, NIH Grant P20 RR016475. Funding to pay the Open Access publication charges for this article was provided by Bioinformatics Core Facility at the University of Kansas.

Conflict of interest statement. None declared.

REFERENCES

- Jones, L.S., Yazzie, B. and Middaugh, C.R. (2004) Polyanions and the proteome. *Mol. Cell. Proteomics*, **3**, 746–769.
- Taylor, J.P., Hardy, J. and Fischbeck, K.H. (2002) Biomedicine— toxic proteins in neurodegenerative disease. *Science*, **296**, 1991–1995.
- Salamat-Miller, N., Fang, J.W., Seidel, C.W., Assenov, Y., Albrecht, M. and Middaugh, C.R. (2007) A network-based analysis of polyanion-binding proteins utilizing human protein arrays. *J. Biol. Chem.*, **282**, 10153–10163.
- Salamat-Miller, N., Fang, J.W., Seidel, C.W., Smalter, A.M., Assenov, Y., Albrecht, M. and Middaugh, C.R. (2006) A network-based analysis of polyanion-binding proteins utilizing yeast protein arrays. *Mol. Cell. Proteomics*, **5**, 2263–2278.
- Conrad, H.E. (1997) *Heparin-Binding Proteins*. Academic Press, New York.
- Lappalainen, P. (2007) *Actin-Monomer-Binding Proteins*. Springer, New York.
- dos Remedios, C.G. and Thomas, D.D. (2000) *Molecular Interactions of Actin: Actin Structure and Actin-Binding Proteins*. Springer, New York.
- Coffman, J.A. and Yuh, C.H. (2004) Identification of sequence-specific DNA binding proteins. *Meth. Cel. Biol.*, **74**, 653–675.
- Chuong, S.D.X., Good, A.G., Taylor, G.J., Freeman, M.C., Moorhead, G.B.G. and Muench, D.G. (2004) Large-scale identification of tubulin-binding proteins provides insight on subcellular trafficking, metabolic channeling, and signaling in plant cells. *Mol. Cell. Proteomics*, **3**, 970–983.
- Killeen, R., Wait, R., Begum, S., Gray, E. and Mulloy, B. (2004) Identification of major heparin-binding proteins in plasma using electrophoresis and mass spectrometry. *Int. J. Exp. Pathol.*, **85**, A69–A69.
- Fang, J.W., Salamat-Miller, N., Dong, Y.H. and Middaugh, C.R. (2007) In Arabnia, H.R. (ed.), *Proceedings of The 2007 International Conference on Bioinformatics & Computational Biology*, Vol. II. CSREA Press, Las Vegas, pp. 427–431.
- Selvaraj, S., Kono, H. and Sarai, A. (2002) Specificity of protein-DNA recognition revealed by structure-based potentials: symmetric/asymmetric and cognate/non-cognate binding. *J. Mol. Biol.*, **322**, 907–915.
- Robison, K., McGuire, A.M., Church, G.M. (1998) A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome. *J. Mol. Biol.*, **284**, 241–254.
- Karmirantzou, M. and Hamodrakas, S.J. (2001) A web-based classification system of DNA-binding protein families. *Protein Eng.*, **14**, 465–472.
- Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M. *et al.* (2006) TRANSFAC (R) and its module TRANSCOMP (R): transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.
- Hung, K.W., Kurnar, T.K.S., Kathir, K.M., Xu, P., Ni, F., Ji, H.H., Chen, M.C., Yang, C.C., Lin, F.P. *et al.* (2005) Solution structure of the ligand binding domain of the fibroblast growth factor receptor: role of heparin in the activation of the receptor. *Biochemistry*, **44**, 15787–15798.
- Murphy, J.W., Cho, Y., Sachpatzidis, A., Fan, C.P., Hodsdon, M.E. and Lolis, E. (2007) Structural and functional basis of CXCL12 (stromal cell-derived factor-1 alpha) binding to heparin. *J. Biol. Chem.*, **282**, 10018–10027.
- Cote, R.G., Jones, P., Apweiler, R. and Hermjakob, H. (2006) The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics*, **7**, 97.