Cell. Mol. Life Sci. 63 (2006) 36–51
1420-682X/06/010036-16
DOI 10.1007/s00018-005-5350-6
© Birkhäuser Verlag, Basel, 2006

Cellular and Molecular Life Sciences

Review

The Membrane Protein Data Bank

P. Raman[b], **V. Cherezov**[a,c] **and M. Caffrey**[a,b,c,*]

[a] College of Science and Materials and Surface Science Institute, University of Limerick, Limerick (Ireland),
Fax: +1 353 61 202568,
e-mail: martin.caffrey@ul.ie
[b] Biophysics Program, The Ohio State University, Columbus, Ohio 43210 (USA)
[c] Chemistry Department, The Ohio State University, Columbus, Ohio 43210 (USA)

Received 3 August 2005; received after revision 18 September 2005; accepted 26 September 2005
Online First 28 November 2005

Dedication. This paper and the Membrane Protein Data Bank celebrate the 20th anniversary of the landmark paper in Nature (1985, **318:** 618–624) describing the first 'high-resolution' three-dimensional structure of a membrane protein, the photosynthetic reaction center from *Rhodopseudomonas (Blastochloris) viridis*.

Abstract. The Membrane Protein Data Bank (MPDB) is an online, searchable, relational database of structural and functional information on integral, anchored and peripheral membrane proteins and peptides. Data originates from the Protein Data Bank and other databases, and from the literature. Structures are based on X-ray and electron diffraction, nuclear magnetic resonance and cryo-electron microscopy. The MPDB is searchable online by protein characteristic, structure determination method, crystallization technique, detergent, temperature, pH, author, etc. Record entries are hyperlinked to the PDB and Pfam for viewing sequence, three-dimensional structure and domain architecture, and for downloading coordinates. Links to PubMed are also provided. The MPDB is updated weekly in parallel with the Protein Data Bank. Statistical analysis of MPDB records can be performed and viewed online. A summary of the statistics as applied to entries in the MPDB is presented. The data suggest conditions appropriate for crystallization trials with novel membrane proteins.

Key words. Channels; lipids; receptors; relational database; transporters

1 Introduction

As cellular gatekeepers, membrane proteins play a vital role in the life of the cell. However, when discussing membrane proteins, a common lament has been that there are too few known structures upon which to base general statements regarding the relationship between structure and function. The situation is steadily improving, and the field must now reckon with a rising number of solved structures. Up to now, when all we had was a handful, the membrane protein structure web sites from the laboratories of H. Michel (http://www.mpibp-frankfurt.mpg.de/michel/public/memprotstruct.html) and S. White (http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html) were extremely useful. However, with the growth in numbers of structures there exists a need for an online, searchable relational database of such data. This has been realized in the Membrane Protein Data Bank (MPDB; http://www.lipidat.chemistry.ohio-state.edu/MPDB/index.asp), a description of which is presented here.

At the heart of the MPDB is an MPDB record, with at least one for each protein or peptide in the database. The

* Corresponding author.

contents of the MPDB record draw heavily on the corresponding Protein Data Bank (PDB, http://www.rcsb.org/pdb/) record [1]. However, the former brings together in a convenient format information of direct interest to the membrane structural and functional biologist. Such information is generally not immediately accessible in the PDB record. The MPDB record is also resourced with a host of direct links to related information (structure, sequence, coordinates, etc.) bearing on the protein of interest.

The PDB is an extraordinarily useful and complete resource. However, its focus is on protein structure regardless of the identity or type of protein. Thus, its current panel of ~31,000 records covers all types from soluble to membrane and structural proteins. To cull information on the membrane protein subset requires that the user perform a search of these records. But, the query tool available at the PDB is not discriminating enough, and a text search under 'membrane' elicits close to 2000 hits. In contrast, however, the current literature refers to structures of a mere 50–60 different membrane proteins. To what, then, can we attribute the disparity? It arises in part from the limited descriptors used by authors to annotate entries in the PDB and to the indiscriminate nature of the search tools provided. The disparity is also due to the limiting criteria used when enumerating membrane proteins of known structure. Thus, for example, the 2000 records referred to above include soluble membrane protein fragments. They also include proteins that interact with membrane proteins but that are not themselves membrane proteins. The MPDB was designed to avoid these shortcomings and to provide a resource specific to membrane proteins and peptides.

Our original intent was to limit the scope of the MPDB to bone fide membrane proteins. These include the classically defined polytopic proteins that cross the membrane several times as well as the bitopic type that cross it just once (fig. 1). In the interests of completeness, the database has been extended to include monotopic and peripheral proteins. The former include proteins which are anchored in but do not cross the membrane. In addition to proteins, the data bank also hosts polypeptides, many of which are antimicrobial agents.

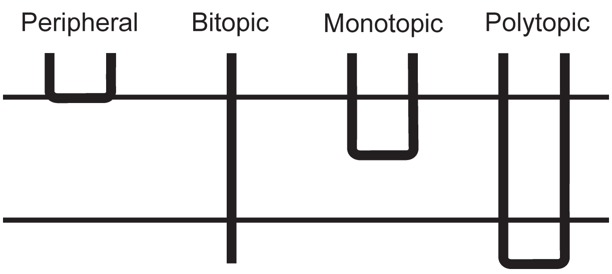


Figure 1. Cartoon representation of the membrane topologies featured in the MPDB. Thin horizontal lines demark the membrane in and on which the protein (bold lines) sits. The simplest polytopic protein, with just two membrane crossings, is shown.

An early version of the MPDB included only structures solved by X-ray crystallography as applied to three-dimensional (3D) crystals. Again, in the interests of completeness the data bank has been extended to structures determined using electron diffraction, cryo-electron microscopy and nuclear magnetic resonance (NMR). The spatial resolution limit for inclusion in the MPDB is 10 Å.

While the scope of the MPDB is thus broad, the fact that it is housed in a relational database means that it can be selectively searched for particular subsets of membrane protein types. It can also be readily updated. The intent is to do so in parallel with the PDB on a weekly basis.

Another attractive feature of the relational nature of the MPDB is that the data therein can be processed and analyzed with ease. Accordingly, the data bank includes a section titled 'Statistics'. These take the form of frequency histograms where a range of parameters, such as crystallization method, detergent, temperature, etc., used for structure determination are shown along with their absolute and relative frequencies of use. The statistics are performed on the most recent update of the database. While not gone into in great detail here, the Statistics feature enables the interested user to mine the MPDB for important trends and relationships. The hope is that this will be used to facilitate more rational and successful structure elucidation strategies and to provide a succinct summary of what has been accomplished in the field and, by extension, what remains to be done.

The paper is organized as follows. We begin with an overview of how the database was constructed and of the criteria used for record inclusion. The database itself is introduced next, beginning with a description of an MPDB record. This leads into an outline of the various search options available for querying the MPDB. The online statistical analysis feature and what the data tell us about conditions conducive to membrane protein and peptide structure determination are described next. The paper ends with a view to future MPDB developments.

2 Methods

2.1 Record source and criteria for inclusion

All records in the MPDB have a corresponding PDB record. An initial search of the PDB under 'membrane protein' produced close to 660 records. These records were evaluated individually, and a decision was made as to their suitability for inclusion in the MPDB based on criteria identified below. Unfortunately, such a search is error prone and can fail to flag bone fide membrane protein records where, for example, the record is identified as referring to a membrane *channel* or *transporter* instead of a membrane *protein*. Accordingly, the search was expanded to include all entries containing 'membrane' and these too were evaluated individually as above.

Once a record was identified as relevant, its header file was downloaded from the PDB. Appropriate data were copied from the header and pasted into the MPDB record. As much as possible, obvious spelling and punctuation errors in the original record were corrected at this stage. Data that were ambiguous or unavailable in the PDB record were retrieved from the source literature and related databases by the MPDB curators and entered into the MPDB record. The Pfam designation and crystallization conditions are two such examples (see section 3.1.2). Occasionally, incomplete data were reported in the PDB record, as, for example, when PEG is identified as the precipitant without reference to its molecular weight and size distribution. If the relevant information was easily available in the source literature it was retrieved and entered in the MPDB record.

As noted, there are extant web sites that contain membrane protein structure data that are based on the PDB. These include the Michel (http://www.mpibp-frankfurt.mpg.de/michel/public/memprotstruct.html), the White (http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html) and the PDB TM (http://pdbtm.enzim.hu/index.php) sites [2]. In the interests of completeness, it was verified that data in the MPDB included all proteins reported in these sites. Our own knowledge of the field, coupled with separate literature searches, was used with a view to being all-inclusive. We do acknowledge, however, that omissions and other errors may have occurred, and we are anxious to have these brought to our attention so that they may be included or corrected in the next MPDB update.

The focus of the MPDB is obviously on membrane proteins. Clearly, all integral membrane proteins are included as are the so-called monotopic proteins. These are proteins that associate with the membrane without crossing it completely. Accordingly, they are anchored in the membrane to varying degrees and include proteins such as prostaglandin H synthase and fatty acid amide hydrolase. There are proteins such as the annexins that also associate with membranes. However, these were not included in the MPDB for the reason that their association is dependent on calcium ions. In other cases, such as with cytochrome *c*, while association with a membrane is possible, interaction appears to be at the interface, with little penetration into the hydrophobic interior of the membrane. While this type of protein has not been included here, version 2 of the MPDB will accommodate such membrane-associating proteins. Cytochromes P450 are another interesting group of proteins. Some are anchored in the membrane. However, to crystallize them the anchor was removed. In cases where this produced a truly water-soluble fragment, the record was not included in the MPDB. When detergents were required to solubilize the protein as part of the crystallization process, the record was included in the database. In certain cases, organic solvents were required to solubilize the protein or peptide. These, too, have been added to the database.

Another difficulty arose in deciding whether to include water-soluble fragments and water-soluble subunits of membrane proteins. The decision was made against inclusion. Thus, neuraminidase and leader peptidase, both of which are integral membrane proteins, have had their extramembranal parts cleaved and the resulting soluble entity crystallized and used in structure determination. These are not in the MPDB. The soluble F_1 domain of the F_1F_0-ATPase (PDB ID: 1BMF) and the cytoplasmic domains (1N9P, 1QRQ) of the potassium channel are other examples of records that, for the same reasons, have not been included either.

While the emphasis is on proteins, polypeptides and protein fragments have also been accommodated in the MPDB. The acceptance criteria include the fact or the suspicion that the peptide is in or strongly associates with the membrane. Accordingly, the pore-forming pentadecapeptide gramicidin is in the MPDB. Magainin and melittin, both amphipathic peptide toxins, have also been included. We have chosen to accept such entries as pertain to fragments of membrane proteins that are membrane-bound such as a transmembranal helix from bovine rhodopsin (1FDF) and glycophorin (1AFO).

Hemolysin (7AHL) is an interesting case in that it starts out life as a water-soluble protein. An active form of the multimeric protein for which there is a 3D structure punches holes in target membranes and is a bone fide integral membrane protein. For this reason, it is in the MPDB.

We have included in the MPDB all records that satisfy the above criteria without regard to resolution (up to a limit of 10 Å) or indeed method used for structure determination. Thus, photosystem II from *Synechococcus elongatus*, which was solved to 3.8 Å resolution and which only shows the C_α coordinates, is included in the interests of completeness. Structures determined using X-ray and electron diffraction, cryo-electron microscopy and nuclear magnetic resonance are part of the MPDB. Of course the method used is embedded in the record. Thus, for those interested in structures based on a particular method or to within a specified resolution range, a suitable search can be conducted to extract the relevant records. While the PDB houses theoretical/predicted membrane protein structures (http://www.rcsb.org/pdb/cgi/models.cgi), these are not in the current version of the MPDB.

2.2 Database architecture

The MPDB has a three-tier architecture. It includes a client tier, an application-server tier and a data-server tier.

The client tier presents data to the user and accepts user input which it then passes to the application-server tier. Since the latter is a web-based application, it takes the form of the web browser, for example, Microsoft Internet Explorer. The application-server tier is where the web server resides, and it incorporates the rules of the application. It is responsible for accepting input from the user, evaluating the input against application logic and passing it on to the data tier. Furthermore, it receives information from the data tier and passes it back to the client tier. The current application uses an Internet Information Services (IIS) server, and rules are implemented in VBScript, JavaScript and ASP. The final data-server tier is responsible for data storage and for accepting and sending information to the application-server tier. Data are currently stored in a Microsoft Access relational database. Given the limited size of the current data set, the database architecture was designed in the interests of rapid retrieval rather than storage capacity. Thus, the MPDB allows fast access and retrieval of records from even complex customized queries because the primary key for most tables is the PDB ID.

3 Results and discussion

3.1 The MPDB

A description of the database, its contents and how to use it follows. Before proceeding however, some common elements in the database will be described.

Banners

The top and bottom of all pages in the MPDB have the following banner:

MAIN PAGE | SEARCH | STATISTICS | USEFUL SITES | CONTACT US

The first three elements return the user to the **Main Page**, the **Search** page and the **Statistics** pages, respectively. This feature avoids the need to use the browser back button repetitively to return to the chosen page from deep within the database. The fourth element is a link to **Useful Sites**. It links to a treasure trove of web-based databases and resources concerned with membranes, proteins, lipids and surfactants. The final element **Contact Us** enables the user to communicate with MPDB staff via e-mail.

For the convenience of the user, across the top and bottom of each MPDB record is the following banner:

PDB SUMMARY | VIEW STRUCTURE | SEQUENCE AND SECONDARY STRUCTURE | MATERIALS AND METHODS | FILE DOWNLOAD/DISPLAY

It provides *direct* links to various resource sites within the corresponding PDB record. These include the record **Summary** page, the **Structure Viewing** page, the **Sequence and Secondary Structure** page, the **Materials and Methods** page, and finally the page where the entire PDB record, with or without coordinates, can be downloaded or displayed.

N/A

Throughout the database the abbreviation 'N/A' is used to denote 'not applicable', 'not available' and 'not any' (none). It is used variously to indicate that information of the type listed is either not applicable to the record of interest or is not available in the corresponding PDB record or literature. Thus, for example, some of the early entries in the PDB failed to indicate the pH at which crystals were grown. In such cases, N/A appears in the MPDB record under pH. 'Not any' is implied when, for example, a structure was solved by XRD, but non-protein components (other than water) are not observed in the model. In this case, all three attributes under Non-Protein Components in Structure are assigned values of N/A.

Not Used

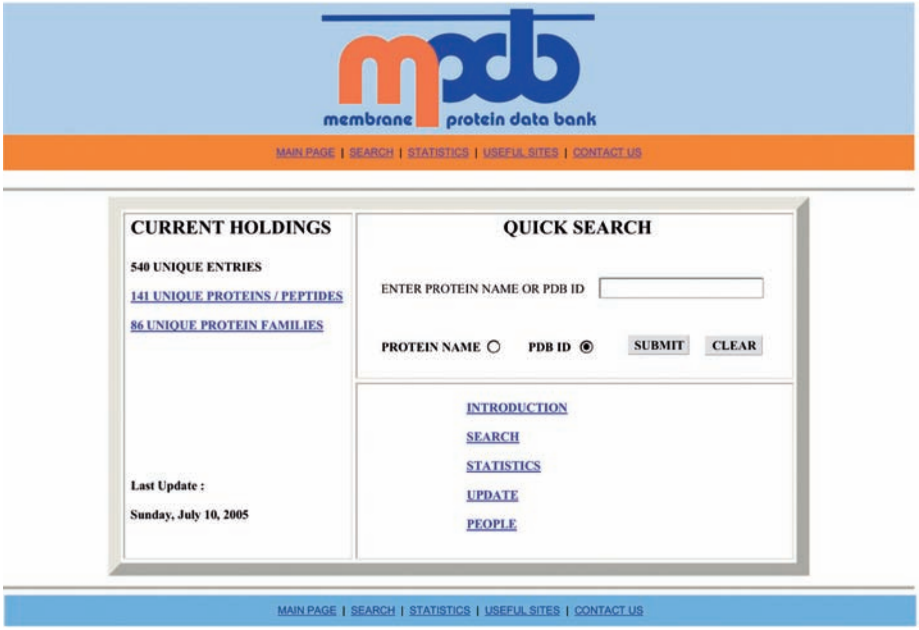
Occasionally, attributes will bear the 'Not Used' descriptor. This caters to situations where the particular attribute was not employed. For example, for certain NMR-based records, lipid was not used in structure determination. In this case, the Lipid Used in Bilayer Crystallization/NMR attribute will have the Not Used descriptor. Another example concerns pH. In situations where a buffer was not employed, the Not Used descriptor is entered under pH.

3.1.1 The Main Page

The MPDB is accessed by way of the **Main Page** (fig. 2) when the following URL, http://www.lipidat.chemistry.ohio-state.edu/MPDB/index.asp, is submitted. It is from the **Main Page** that the user can access the **Search** options page or proceed to examine a host of facts and figures about records in the database by way of the **Statistics** link. Alternatively, the **Quick Search** option can be used directly. This enables the user to enter the name (or the first few letters of the name or a synonym) of the protein of interest or a PDB ID, having activated the appropriate radio button, and to find relevant records in the MPDB. This generates a summary table from which the user chooses the record of interest.

The **Main Page** also provides summary information about current holdings in the MPDB. Listed here are the total number of records in the MPDB, and the total number of records for unique proteins or peptides and families. 'Current Holdings' values are updated automatically on a weekly basis.

Separate links are included on the **Main Page** that take the user to a brief introduction to the MPDB and to the individuals associated with the creation and maintenance of the data bank.


membrane protein data bank
MAIN PAGE | SEARCH | STATISTICS | USEFUL SITES | CONTACT US
SUBMIT
CLEAR
INTRODUCTION
SEARCH
STATISTICS
UPDATE
PEOPLE
MAIN PAGE | SEARCH | STATISTICS | USEFUL SITES | CONTACT US

CURRENT HOLDINGS

540 UNIQUE ENTRIES

141 UNIQUE PROTEINS / PEPTIDES

86 UNIQUE PROTEIN FAMILIES

Last Update :

Sunday, July 10, 2005

QUICK SEARCH

ENTER PROTEIN NAME OR PDB ID

PROTEIN NAME ○ PDB ID ●

Figure 2. The MPDB **Main Page**.

3.1.2 The MPDB record

All searches ultimately lead to an MPDB record (fig. 3). Currently, there are 539 unique record entries in the data bank. A fraction of the information in each was obtained automatically, by means of a script, from the source record in the PDB.

Each record is headed by a title taken verbatim from the PDB record. Below this is the PDB identification code, which is hyperlinked to the corresponding record in the PDB. The remainder of the MPDB record is divided into the following sections.

3.1.2.1 Protein Characteristics

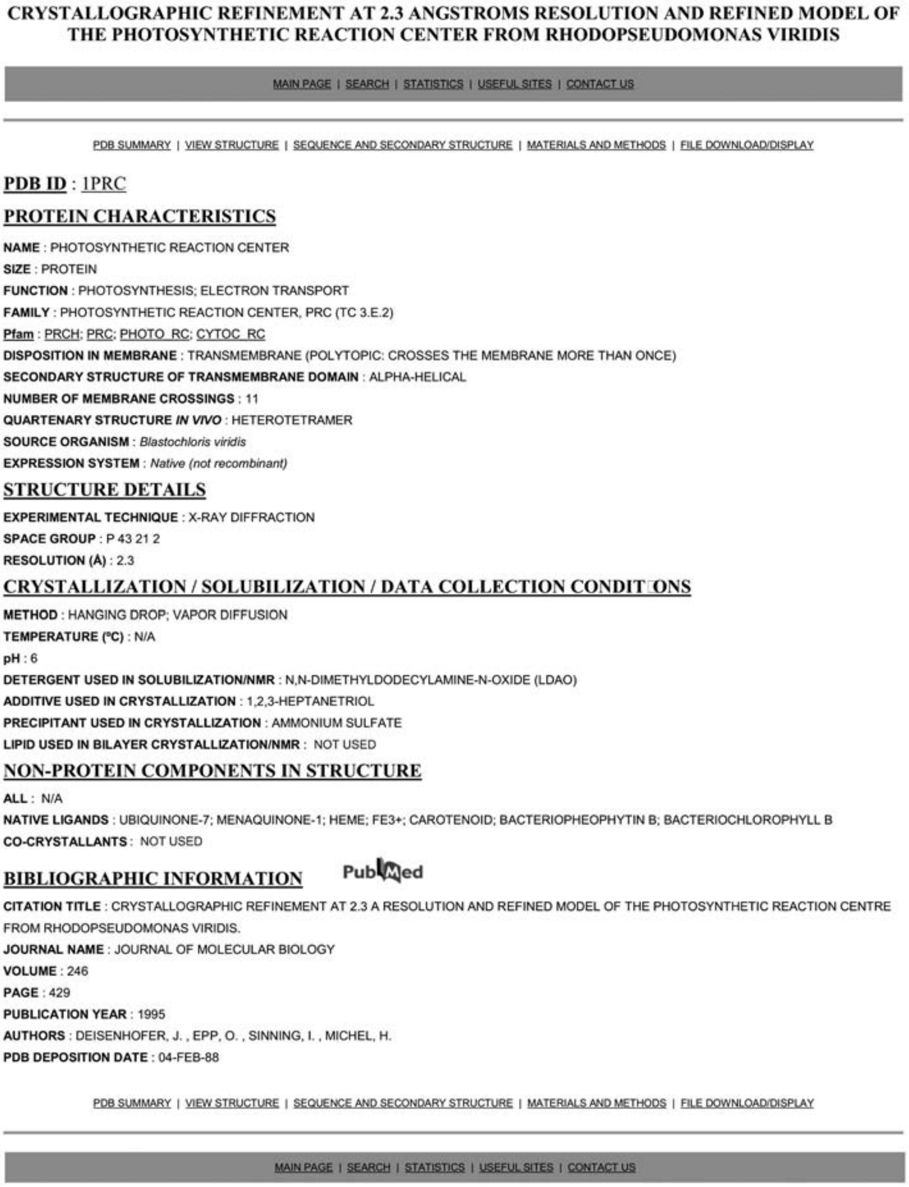
3.1.2.1.a Name This is the name of the protein associated with the corresponding structure record in the PDB. Unfortunately, there is little consistency in the protein-naming process, and the list includes a mix of common names and more formal designations.

3.1.2.1.b Size Proteins and peptides in the database have been categorized based on size, which roughly refers to number of amino acid residues. The terminology used follows the IUPAC Gold Book recommendations (http://www.chemsoc.org/chembytes/goldbook/). Peptides with fewer than 10 residues are designated oligopeptides (no entries as of this writing), while polypeptides are 10–100 amino acids long. Generally, proteins are designated as having in excess of 100 residues as a monomer or as a multisubunit complex. We have included under Size an additional category called Protein Fragment. This refers to proteins that have had a sizable part of their total mass removed or that do not have a full complement of subunits, to facilitate crystallization or NMR structure determination, for example. OmpA (1QJP) and NalP (1UYN) were both crystallized as fragments that retained the transmembrane part only. The two-subunit complex of cytochrome *c* oxidase (1AR1) is another example of what we have classified as a fragment.

3.1.2.1.c Function The functional activity of the protein is indicated here and was obtained directly from the PDB and/or from the literature.

3.1.2.1.d Family Family designation is based on the International Union of Biochemistry and Molecular Biology (IUBMB) classification for enzymes and transporters. Where Enzyme Commission (EC) and Transporter Classification (TC) identifiers were available, they have been included in the entry. When such designations were not available, a family description has been provided by the database curators based on the name and function of the protein.

3.1.2.1.e Pfam Pfam (http://www.sanger.ac.uk/Software/Pfam/index.shtml) is a database of protein domain clans and families. Proteins can consist of one or several domains many of which have been identified and assigned a Pfam code. Codes associated with a sizable fraction of the records in the MPDB (483 out of 539) were found by querying the Pfam database based on the corresponding PDB ID or the protein amino acid sequence and were entered under Pfam in the MPDB record. They are hyperlinked back to the Pfam database, where a complete description of the domain and its interactions and complexes is given.



CRYSTALLOGRAPHIC REFINEMENT AT 2.3 ANGSTROMS RESOLUTION AND REFINED MODEL OF THE PHOTOSYNTHETIC REACTION CENTER FROM RHODOPSEUDOMONAS VIRIDIS

MAIN PAGE | SEARCH | STATISTICS | USEFUL SITES | CONTACT US

PDB SUMMARY | VIEW STRUCTURE | SEQUENCE AND SECONDARY STRUCTURE | MATERIALS AND METHODS | FILE DOWNLOAD/DISPLAY

PDB ID : 1PRC

PROTEIN CHARACTERISTICS

NAME : PHOTOSYNTHETIC REACTION CENTER
SIZE : PROTEIN
FUNCTION : PHOTOSYNTHESIS; ELECTRON TRANSPORT
FAMILY : PHOTOSYNTHETIC REACTION CENTER, PRC (TC 3.E.2)
Pfam : PRCH; PRC; PHOTO_RC; CYTOC_RC
DISPOSITION IN MEMBRANE : TRANSMEMBRANE (POLYTOPIC: CROSSES THE MEMBRANE MORE THAN ONCE)
SECONDARY STRUCTURE OF TRANSMEMBRANE DOMAIN : ALPHA-HELICAL
NUMBER OF MEMBRANE CROSSINGS : 11
QUARTENARY STRUCTURE *IN VIVO* : HETEROTETRAMER
SOURCE ORGANISM : *Blastochloris viridis*
EXPRESSION SYSTEM : *Native (not recombinant)*

STRUCTURE DETAILS

EXPERIMENTAL TECHNIQUE : X-RAY DIFFRACTION
SPACE GROUP : P 43 21 2
RESOLUTION (Å) : 2.3

CRYSTALLIZATION / SOLUBILIZATION / DATA COLLECTION CONDIT□ONS

METHOD : HANGING DROP; VAPOR DIFFUSION
TEMPERATURE (ºC) : N/A
pH : 6
DETERGENT USED IN SOLUBILIZATION/NMR : N,N-DIMETHYLDODECYLAMINE-N-OXIDE (LDAO)
ADDITIVE USED IN CRYSTALLIZATION : 1,2,3-HEPTANETRIOL
PRECIPITANT USED IN CRYSTALLIZATION : AMMONIUM SULFATE
LIPID USED IN BILAYER CRYSTALLIZATION/NMR : NOT USED

NON-PROTEIN COMPONENTS IN STRUCTURE

ALL : N/A
NATIVE LIGANDS : UBIQUINONE-7; MENAQUINONE-1; HEME; FE3+; CAROTENOID; BACTERIOPHEOPHYTIN B; BACTERIOCHLOROPHYLL B
CO-CRYSTALLANTS : NOT USED

BIBLIOGRAPHIC INFORMATION PubMed

CITATION TITLE : CRYSTALLOGRAPHIC REFINEMENT AT 2.3 A RESOLUTION AND REFINED MODEL OF THE PHOTOSYNTHETIC REACTION CENTRE FROM RHODOPSEUDOMONAS VIRIDIS.
JOURNAL NAME : JOURNAL OF MOLECULAR BIOLOGY
VOLUME : 246
PAGE : 429
PUBLICATION YEAR : 1995
AUTHORS : DEISENHOFER, J. , EPP, O. , SINNING, I. , MICHEL, H.
PDB DEPOSITION DATE : 04-FEB-88

PDB SUMMARY | VIEW STRUCTURE | SEQUENCE AND SECONDARY STRUCTURE | MATERIALS AND METHODS | FILE DOWNLOAD/DISPLAY

MAIN PAGE | SEARCH | STATISTICS | USEFUL SITES | CONTACT US

Figure 3. An MPDB record.

3.1.2.1.f Disposition in Membrane Disposition refers to the topology of the protein with respect to the membrane in or on which it sits. Four membrane disposition types have been created (fig. 1). The first is designated anchored or monotopic, where the protein is embedded in but does not cross the membrane. The second is of the peripheral type, where the protein is loosely associated with the membrane. Cytochrome P450, β-glucosidase (1VFF) and the juxtamembrane domain of the epidermal growth factor receptor (1Z9I) are currently the only entries in this group. The third category includes transmembranal proteins of the bitopic type which cross the membrane just once. This group includes a host of peptides and protein fragments, and it consists mostly of NMR structures. Transmembranal proteins of the polytopic type that cross the membrane more than once form the last and most populous of all the categories.

3.1.2.1.g Secondary Structure of Transmembranal Domain This characteristic refers to the dominant, secondary structure of the membrane-crossing part of the protein. The two major categories are α-helical and β-sheet, the latter typically in the form of a β-barrel. Helical-barrel is a third category that accommodates the several gramicidin records. And finally there is the N/A category, which covers peripheral and monotopic proteins.

3.1.2.1.h Number of Membrane Crossings Since this is an important characteristic of integral membrane proteins, the relevant data have been enumerated by us and annotated for each transmembranal protein in the data bank. The number of crossings, as helices or strands, is reported on a per monomer basis for homomultimeric proteins, such as LH2. In the case of heteromultimers, cytochrome c oxidase, for example, the number of crossings is for the entire complex. Other illustrative examples include the homotetramer aquaporin 1 (1J4N). Each aquaporin monomer has 6 transmembrane helices and 2 membrane-embedded helices. Accordingly, the reported number for this entry is 6. The ClC channel (1KPL) is a homodimer. Each monomer has totally 18 α-helices, but only 10 span the membrane. The reported number is 10.

Occasionally, a membrane crossing occurs by means of a long helix with a short break. In the MPDB, this is counted as a single crossing. The crossing extending from residue 121 to 143 in the Na^+/H^+ antiporter NhaA (1ZCD) is one such example.

3.1.2.1.i Quaternary Structure in Vivo The native quaternary structure of a protein, if known, is entered here. Examples include homodimers and heteropentamers as in the case of the ClC chloride channel (1KPL) and the nicotinic acetylcholine receptor (2BG9), respectively.

3.1.2.1.j Source Organism This section contains the Latin name of the species of organism from which the protein or peptide was obtained directly, or by means of homo- or heterologous expression. The list also includes viruses.

The sequence of synthetic and recombinant proteins and peptides used for structure determination is usually based on that of a natural homolog. In this case, the relevant source is identified under Source Organism.

3.1.2.1.k Expression System The bulk of the entries in the MPDB refer to proteins produced using recombinant DNA technology. This requires a homo- or heterologous expression system of which there are 16 types represented in the MPDB, as of this writing. The Latin name of the organism in which the expression was done is identified here. One of the options within this category is identified as ‘Native’. This refers to proteins that were produced in the native organism and which are not recombinant proteins. The list includes ‘Synthetic’, which refers to ‘man-made’ peptides whose sequence is based on that of a natural protein or peptide.

3.1.2.2 Structure Details

3.1.2.2.a Experimental Technique The method used for structure determination is identified here. The methods available include cryo-electron microscopy, electron diffraction, nuclear magnetic resonance (NMR) and X-ray diffraction (XRD).

3.1.2.2.b Space Group This feature is relevant only to structures solved by means of diffraction. The value is obtained directly from the corresponding PDB record.

3.1.2.2.c Resolution The spatial resolution with which the structure was solved is provided here. Values are reported for all entries with the exception of those based on NMR, for which the N/A designation is applied.

3.1.2.3 Crystallization / Solubilization / Data Collection Conditions

3.1.2.3.a Method This refers to the procedure by which crystals were grown for use in structure determination. The vast majority of the information entered here was obtained by us from the literature since it is not always found in the PDB, especially in older records. The descriptors employed by the authors of the source articles are used. Obviously, for structures determined by non-crystallographic means, the N/A descriptor is used.

3.1.2.3.b Temperature For structures solved by crystallographic means, this refers to the temperature at which crystallogenesis was performed. For NMR-based structures, it is the temperature at which NMR data were collected.

3.1.2.3.c pH For structures solved by crystallographic means, this refers to the pH at which crystallogenesis was performed. For NMR-based structures, it is the pH of the solution used for NMR data collection.

3.1.2.3.d Detergent Used in Solubilizationn/NMR Detergents are invariably used at one or several stages during structure determination for purposes of protein solubilization from the source membrane or inclusion body. Thus, the detergent identified is the one employed in crystallogenesis when applied to XRD structures. In the case of NMR structures, it generally refers to the detergent used to solubilize the peptide or protein for solution NMR data collection.

3.1.2.3.e Additive Used in Crystallization Crystallization protocols typically require the inclusion of what are loosely referred to as ‘additives’ that are often small amphiphiles. Ethylene glycol, heptanetriol, EDTA, dithiothreitol, as well as small amounts of NaCl are examples of typical additives. D_2O, used in the crystallization of BtuCD (1L7V), is another example.

3.1.2.3.f Precipitant Used in Crystallization The additive included in the crystallization mix for purposes of inducing nucleation and crystal growth is referred to as the precipitant. Ammonium sulfate and the polyethylene glycols (PEGs) are common precipitants.

3.1.2.3.g Lipid Used in Bilayer Crystallization/NMR For XRD records, this covers lipids used in crystallization by the so-called bilayer (cubic phase, vesicle fusion, bicelle) methods [reviewed in ref. 3]. To date, there are only four lipid species that have been used for this purpose. In the case of NMR entries, lipids are used to form bicelles and oriented bilayers with which to collect structure information.

3.1.2.4 Non-Protein Components in Structure

3.1.2.4.a All A listing of all of the non-proteinaceous materials, other than water, that appear in the final structure are presented here. These can include ions, additives, precipitants, lipids, detergents, cofactors, prosthetic groups and pigments, among other things. The relevant data with which this field was populated was obtained directly from the 'Heterogen' entry in the corresponding PDB record.

3.1.2.4.b Native Ligands This is a subset of the components listed in section 3.1.2.4.a that are a part of the native protein. Native ligands include pigments, prosthetic groups, ions and the like.

3.1.2.4.c Co-crystallants A co-crystallant is an ion, a small molecule or macromolecule (antibody fragment, for example) that was combined with the target protein for purposes of crystallizing it as a complex.

3.1.2.5 Bibliographic Information

3.1.2.5.a Citation Title This is the title of the primary citation as it appears in the PDB. Note that the primary citation may refer to a paper in which a refined version of the original model was reported.

3.1.2.5.b Journal Name/Volume/Page/Publication Year/Authors The primary citation (journal reference) to the structure is listed here. The journal volume, start page number and year of publication, along with a complete listing of authors are included in the MPDB record. One of the records in the MPDB is based on a thesis and is so identified under Journal.

A direct link to the corresponding publication in PubMed Central, an archive of life sciences journals (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=Pubmed), is available by way of the PubMed icon to the right of the 'Bibliographic Information' heading. At the very least, an abstract of the article is posted there. In many cases, a full text pdf of the paper is also available for online viewing and/or download.

3.1.2.5.c PDB Deposition Date This refers to the date on which the structure coordinates were deposited in the PDB. It is important to note that refinements and corrections to records in the PDB are possible. When major changes are made, a new PDB ID is assigned which replaces the extant ID. The deposition date and the date of publication of the primary citation usually coincide closely. However, in some early PDB records a significant mismatch exists. The photosynthetic reaction center from the *Rhodopseudomonas (Blastochloris) viridis* record (1PRC) is a case in point. It has a deposition date of 1988. Several revisions were made to the record between 1989 and 1994. Publications cited date back to 1982, when first crystals were reported [4]. The original structure at 3 Å resolution was published in 1985 [5], and in 1995 a refined model was described [6]. 1995 is the publication year of the primary citation posted in the PDB and the MPDB.

3.2 Searching the MPDB

One of the primary reasons for creating the MPDB was to provide a convenient interface with which the contents of the MPDB could be perused and interrogated with efficiency and ease. As noted, a somewhat limited but immediate 'Quick Search' option is available from the **Main Page**. However, more complete and extensive searches can be conducted via the **Search** page, which is accessed from the **Main Page** by activating the **Search** links (fig. 2).

The **Search** page has a list of alphabetically arranged options by means of which the MPDB can be queried. One of these, the PDB ID option, takes the user directly to an MPDB record. All others generate a table of records (summary table, fig. 4) the entries of which satisfy the search criteria. The nature of the search and the number of records found are shown above and to the right hand side of the summary table. The table includes the PDB ID, title and spatial resolution associated with each record. Entries in the table are arranged by resolution. The user can go to the record of choice by clicking on the appropriate table entry. When the total number of records in a table exceeds 10, the remaining entries can be viewed on separate pages that are numbered to the bottom left of each table. The page number of the displayed page is shown in bold and black.

Most queries are presented in the form of dropdown lists. List entries are arranged alphabetically for easy searching.

The search options available in the MPDB are described individually below.

3.2.1 Author

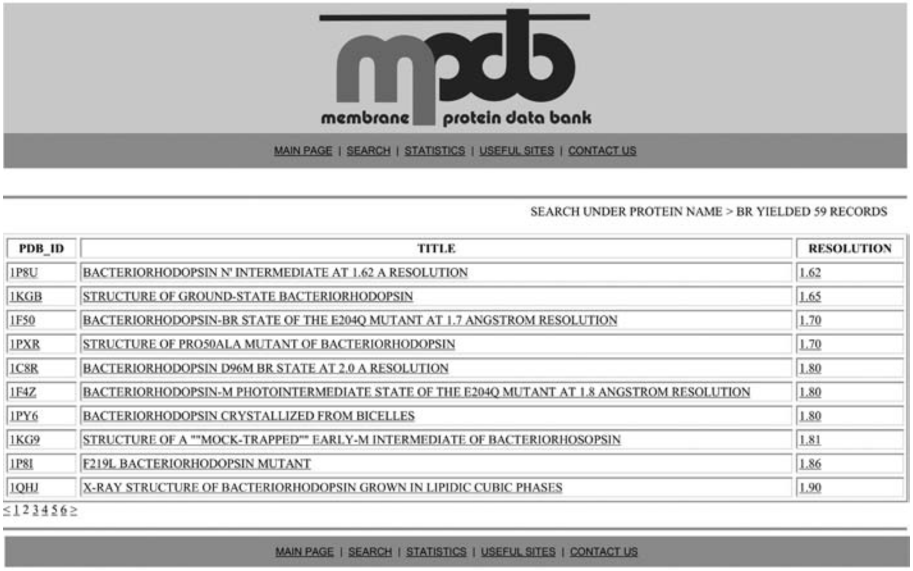
A dropdown list of authors, alphabetized by last name, is available under the **Author** search option.

3.2.2 Co-crystallant

The basis for this option has been described above (section 3.1.2.4.c).

3.2.3 Crystallization Method

The dropdown list for this query includes the following options: antibody, batch, bicelle, dialysis, epitaxial nucleation, hanging drop, lipid cubic phase, macroseeding,


membrane protein data bank
MAIN PAGE | SEARCH | STATISTICS | USEFUL SITES | CONTACT US

SEARCH UNDER PROTEIN NAME > BR YIELDED 59 RECORDS

PDB_ID	TITLE	RESOLUTION
1P8U	BACTERIORHODOPSIN N' INTERMEDIATE AT 1.62 A RESOLUTION	1.62
1KGB	STRUCTURE OF GROUND-STATE BACTERIORHODOPSIN	1.65
1F50	BACTERIORHODOPSIN-BR STATE OF THE E204Q MUTANT AT 1.7 ANGSTROM RESOLUTION	1.70
1PXR	STRUCTURE OF PRO50ALA MUTANT OF BACTERIORHODOPSIN	1.70
1C8R	BACTERIORHODOPSIN D96M BR STATE AT 2.0 A RESOLUTION	1.80
1F4Z	BACTERIORHODOPSIN-M PHOTOINTERMEDIATE STATE OF THE E204Q MUTANT AT 1.8 ANGSTROM RESOLUTION	1.80
1PY6	BACTERIORHODOPSIN CRYSTALLIZED FROM BICELLES	1.80
1KG9	STRUCTURE OF A ""MOCK-TRAPPED"" EARLY-M INTERMEDIATE OF BACTERIORHOSOPSIN	1.81
1P8I	F219L BACTERIORHODOPSIN MUTANT	1.86
1QHJ	X-RAY STRUCTURE OF BACTERIORHODOPSIN GROWN IN LIPIDIC CUBIC PHASES	1.90

≤ 1 2 3 4 5 6 ≥

Figure 4. A typical MPDB summary table.

microseeding, sitting drop, under oil, vapor diffusion and vesicle fusion. These are the descriptors used by authors to identify the methods used for crystallization. The methods are not mutually exclusive. Thus, for example, the sitting drop and hanging drop methods are forms of vapor diffusion. And 'batch' includes 'under oil' and 'lipid cubic phase' methods. At first, we had intended to devise and to implement a scheme that would properly describe these assorted crystallization methods and how they are related. But we felt that the new notation that would have to be introduced would not be familiar and would impede efficient use of the database. Such a rationalization of crystallization methods is badly needed. It will not be addressed further here.

3.2.4 Crystallization/Solubilization/Data Collection Conditions

This option enables the user to search the database based on one or several criteria in combination. There are four such criteria. The first concerns the detergent used in solubilizing the protein. This covers diffraction, microscopy and NMR-based records, since detergents are generally used at some point during structure determination by virtually all methods. The second and third criteria refer to additives and precipitants employed in crystallization. The fourth covers lipids used in crystallization by the bilayer methods. For each of the four criteria, the user selects from a dropdown list.

To perform a search based on just one of the four criteria, the user simply clicks on the dropdown list of choice, highlights a value on the list and submits the query. All other criteria windows are left unaltered, displaying the default value 'ALL'. The latter indicates that the data bank is being searched under the selected criteria with all others unspecified and, thus, open to ALL values. The **Customized Search**, described next, uses the ALL feature extensively.

3.2.5 Customized Search

The **Customized Search** was created to enable the user to create and to run a query based on a combination of provided search keys. This is in contrast to the bulk of the other **Search** options that are based on just one or, at most, a few select keys. The user clicks on the appropriate boxes and chooses from dropdown lists to specify the query. All other options within the **Customized Search** are left in their default settings. The search is run and the results are presented in tabular form. The record of choice is selected from within the summary table by point and click.

3.2.6 Disposition in Membrane

This option enables the user to search the database with respect to the topology of the protein in or on the membrane based on four possible choices.

3.2.7 Experimental Technique

It is here that the user gets to choose one of four methods by which structure was determined as follows: cryo-electron microscopy, electron diffraction, NMR and XRD.

3.2.8 Expression System

The basis for this option was described above (section 3.1.2.1.k).

3.2.9 Family

The basis for this option has been described (section 3.1.2.1.d).

3.2.10 Function

The basis for this option has been described (section *3.1.2.1.c*).

3.2.11 Journal

The journal publishing the article in which the PDB record originated is listed here. One of the entries in the MPDB is based on a thesis and is identified as such.

3.2.12 Name

This refers to the name assigned to the protein by the authors. But many proteins have formal as well as common names and abbreviations. As much as possible, all synonyms are included in the **Search by Name** dropdown list to facilitate finding the protein of interest. Thus, for example, the Vitamin B12 receptor can be located under 'Vitamin B12 receptor', 'BtuB', and 'outer membrane cobalamin transporter'. There is also the option of entering the protein name, or part thereof, into a text dialog box. To use this option the default setting 'None' in the dropdown list must be selected.

3.2.13 Native Ligand

Many of the proteins listed in the MPDB contain natural pigments, cofactors and prosthetic groups. These can be searched for via the dropdown list identified under Native Ligand.

3.2.14 Non-Protein Component in Structure

This refers to components other than water that appear in the final structure and that are not naturally a part of the protein. They include ions, detergents, lipids, cofactors, among others. By and large, such non-proteinaceous materials derive from the solutions used to grow the crystal upon which the structure is based. Some of the components have long and complex names. To make the dropdown list fit the screen, some of these names have been abbreviated. In such cases, the user should consult the PDB record for the full name.

3.2.15 PDB ID

This is the identification code associated with the corresponding record in the PDB. Records can be selected from a dropdown list or by entering text into a dialog box.

3.2.16 pH

In the case of records that are based on XRD data, this refers to the pH at which crystals were grown. For NMR-based records, it is the pH of the solution used for data collection. A particular pH value or a range of pH values can be used to perform a search. Radio buttons enable the user to have included in the summary table records for which a pH value was not specified. In addition, the experimental technique used for structure determination can be selected so as to limit the search.

3.2.17 Publication Year

The year in which the primary citation reporting the particular PDB record was published can be used as a search criterion. (It is important to note that the Publication Year can differ from that when the PDB entry was made. In most cases, the latter PDB Deposition Date predates but is close in time to the year of publication. This is particularly true of late now that deposition in the PDB is generally a prerequisite for publication and refinements are not common. However, in the example cited above for the photosynthetic reaction center (section 3.1.2.5.c), the year of publication for the *primary citation* in the PDB is 7 years post-deposition.) In addition to being able to specify a particular year, it is also possible to use this option to search over a selectable period from 1982 to 2005.
Some fraction of records in the MPDB reside in a category titled 'To Be Published'. These represent structures that have been posted in the PDB, but for whatever reason, the authors have not gotten round to reporting in the literature. Radio buttons are provided so that the 'To Be Published' records are included or excluded from the summary table.

3.2.18 Resolution

The resolution with which a structure has been determined to is criterion upon which a search can be performed. The posted resolution values extend from 0 to 10 Å, and specific values, as well as ranges, can be searched. Resolution is reported for all techniques with the exception of NMR, in which case the resolution value entered is N/A. Radio buttons are provided that enable the user to include or exclude such records from the search results.

3.2.19 Secondary Structure of Transmembrane Domain

This query enables the user to perform a search based on the dominant secondary structure of the membrane crossing part of the protein, as described above (section 3.1.2.1.g). For proteins with α-helical or β-barrel transmembrane domains, the search can be performed by specifying the number of membrane crossings.

3.2.20 Size

The size of a protein or peptide, or fragment thereof, as defined above (section 3.1.2.1.b) can be searched for under this attribute.

3.2.21 Source Organism

The source organism refers to the species of organism from which the protein was derived directly or by means of homo- or heterologous expression.

3.2.22 Temperature

This is the temperature at which crystallogenesis was performed or, in the case of NMR-based records, the temperature at which spectra were recorded. A range of temperatures can be selected in addition to a single temperature. For several entries in the database an appropriate temperature could not be found in the PDB record and corresponding literature. In this case, the temperature value assigned is N/A. By means of radio buttons the user can choose to include or exclude such records in a given search. It is also possible to refine the temperature search by specifying the Experimental Technique.

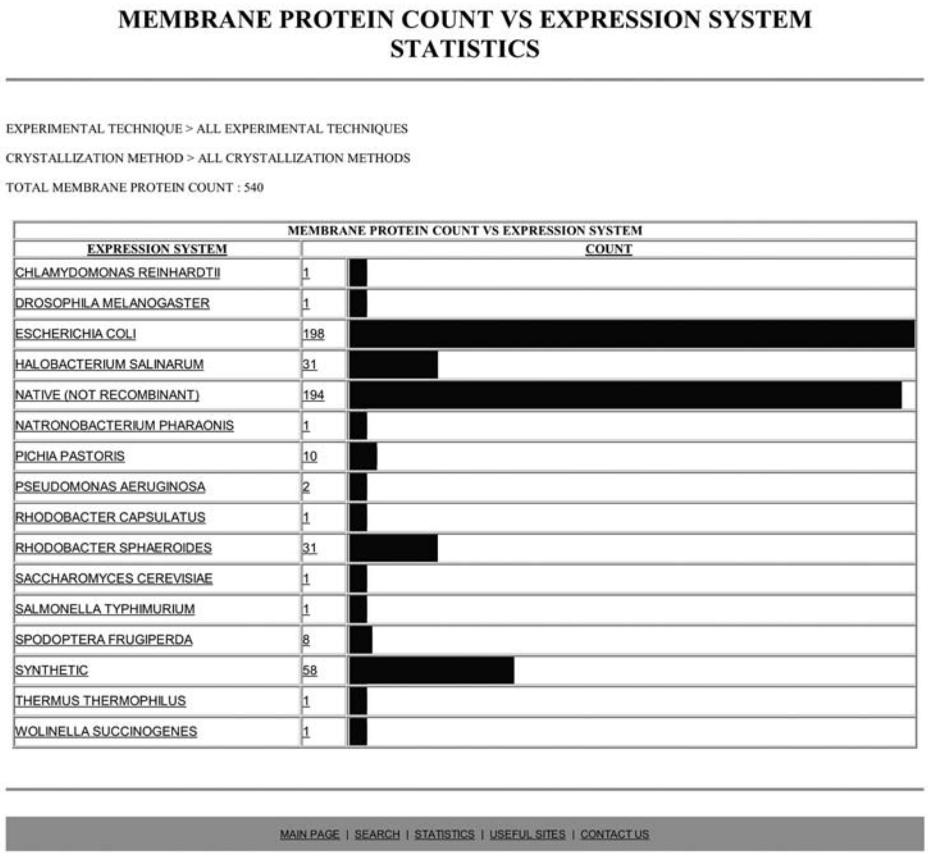
3.3 Facts and figures

An assortment of statistics can be generated online based on the contents of the MPDB. These can be accessed from the **Main Page** by way of the **Statistics** link (fig. 2). The user can choose the statistical data of interest from a dropdown list which, in turn, generates a table and a graphical representation of the data (fig. 5). The plot takes the form of a histogram oriented horizontally to facilitate labeling and to make the independent variable entries more legible. The element with the highest record count is assigned the maximum bar length on a scale that is linear. Clicking on **Count** above the histogram enables the data to be sorted in order of increasing record count from top to bottom. A second click inverts the sort. The same feature applies to the heading above the table on the left. Clicking on it arranges table entries alphabetically and/or numerically. Statistics are calculated in SQL based on the latest weekly update. Those records that are included in a given count can be viewed by clicking on the given element or count in the table.

Additional statistical analysis can be added in future versions of the MPDB. Users are encouraged to bring these to our attention.

From the **Statistics** page the user can perform an analysis on the entire contents of the MPDB by simply clicking on the **Pick a Statistic** button and selecting from the dropdown list. In this case, the default settings are retained for the other two features on the **Statistics** page. However, it is also possible to limit the scope of the analysis to one and/or two subsets of the database. In this case, the type of analysis is selected as above and the subset upon which the analysis is to be performed is chosen by selecting from the **Experimental Technique** and/or **Crystallization Method** dropdown lists. Thus, for example, it is possible to perform an analysis of 'number of records versus year' over all entries in the MPDB. Or, by selecting the appropriate subsets, 'annual record frequency' can be obtained for those structures determined using 'XRD' and the 'hanging drop method'.

In what follows, we describe the statistical analyses options that are currently available in the MPDB and later discuss them in terms of how the data might inform future membrane protein crystallization endeavors.



MEMBRANE PROTEIN COUNT VS EXPRESSION SYSTEM STATISTICS

EXPERIMENTAL TECHNIQUE > ALL EXPERIMENTAL TECHNIQUES

CRYSTALLIZATION METHOD > ALL CRYSTALLIZATION METHODS

TOTAL MEMBRANE PROTEIN COUNT : 540

MEMBRANE PROTEIN COUNT VS EXPRESSION SYSTEM

EXPRESSION SYSTEM	COUNT
CHLAMYDOMONAS REINHARDTII	1
DROSOPHILA MELANOGASTER	1
ESCHERICHIA COLI	198
HALOBACTERIUM SALINARUM	31
NATIVE (NOT RECOMBINANT)	194
NATRONOBACTERIUM PHARAONIS	1
PICHIA PASTORIS	10
PSEUDOMONAS AERUGINOSA	2
RHODOBACTER CAPSULATUS	1
RHODOBACTER SPHAEROIDES	31
SACCHAROMYCES CEREVISIAE	1
SALMONELLA TYPHIMURIUM	1
SPODOPTERA FRUGIPERDA	8
SYNTHETIC	58
THERMUS THERMOPHILUS	1
WOLINELLA SUCCINOGENES	1

Figure 5. A screen shot of a default statistical analysis performed on all MPDB records versus expression system.

3.3.1 Additive Used in Crystallization

To date, 63 different additives have been used to facilitate the crystallization of membrane proteins whose structures are reported in the MPDB. As noted, additives are generally small amphiphiles that support crystallogenesis. But since the definition is not strict and no convention has been accepted in regard to use of the term 'additive', authors use the word with some abandon. As a result, the list of additives includes everything from small amphiphiles, to lipids, polymers and salts. Indeed, the most frequently encountered additive in the MPDB set is sodium chloride followed closely by 1,2,3-heptanetriol. The next in order of popularity is glycerol, sodium azide (an antimicrobial agent), 2-methyl-2,4-pentanediol (MPD) and magnesium chloride (table 1).

3.3.2 Crystallization Method

As explained above (section 3.1.2.3.a), the methods used for crystallizing membrane proteins with representation in the MPDB have been divided into 13 categories that

Table 1. Conditions, methods and systems used in membrane protein and peptide structure determination and their contribution to record count in the MPD (the table includes the top 10 entries in each category arranged by record count).

Conditions, methods, systems	Count
Additive[1]	
Sodium chloride	58
1,2,3-Heptanetriol	56
Glycerol	32
Sodium azide	22
2-Methyl-2,4-pentanediol	20
Magnesium chloride	19
EDTA	15
Dithiothreitol	14
1,6-Hexanediol	13
Calcium chloride	13
Detergent[1]	
N-Octyl-ß-d-glucopyranoside (OG)	97
N,n-Dimethyldodecylamine-n-oxide (LDAO)	77
Octyltetraoxyethylene (C8E4)	44
N-Dodecyl-ß-d-maltopyranoside	38
N-Decyl-ß-d-maltopyranoside	28
Dodecylnonaoxyethylene (C12E9)	20
N-Octyl-2-hydroxyethylsulfoxide	15
Octylpolyoxyethylene	14
CHAPSO	13
N-Nonyl-ß-d-glucopyranoside	11
Precipitant[1]	
Polyethylene glycol 4000	84
Polyethylene glycol 400	39
Ammonium sulfate	35
Polyethylene glycol 2000	35
Sodium/potassium phosphate	27
Potassium phosphate	25
Polyethylene glycol 600	23
Polyethylene glycol monomethylether 2000	16
Magnesium chloride	15
2-Methyl-2,4-pentanediol	14
Temperature (°C)[1]	
20	96
4	63
18	35
25	30
22	27
19	15
23	15
37	12
16	7
5	7
Crystallization Method[1,2]	
Vapor diffusion	324
Hanging drop	145
Sitting drop	125
Batch	53
Lipid cubic phase	38
Dialysis	32
Antibody	18
Bicelle	13
Microseeding	7
Vesicle fusion	7
Expression System[3]	
Escherichia coli	198
Synthetic	58
Halobacterium salinarum	31
Rhodobacter sphaeroides	31
Pichia pastoris	10
Spodoptera frugiperda	8
Pseudomonas aeruginosa	2
Wolinella succinogenes	1
Thermus thermophilus	1
Salmonella typhimuriam	1
Source organism[3]	
Escherichia coli	115
Halobacterium salinarum	60
Rhodobacter sphaeroides	49
Homo sapiens	48
Bos taurus	34
Bacillus brevis	19
Ovis aries	17
Alicyclobacillus acidocaldarius	16
Oryctolagus cuniculus	15
Streptomyces lividans	12

[1] Refers to XRD records.
[2] Not all methods are mutually exclusive.
[3] Refers to all records.

are not mutually exclusive. Thus, there is overlap between certain categories. To run the statistics on **Crystallization Method**, the only really useful option is to perform it under default conditions. This generates a table and a plot showing record count versus method. A summary of records within a methods category can be obtained simply by clicking on the method or the count within the table.

Of the crystallization methods available, vapor diffusion has been by far the most extensively used (table 1). Within this group, hanging drop with 145 records is ahead of the sitting drop approach with 125 records. Batch, followed successively by the lipidic cubic phase, dialysis, antibody and bicelle methods round out the other major contributors. The epitaxial nucleation, micro- and macroseeding, vesicle fusion and batch- under-oil methods have all contributed to the total yield of membrane protein structures but in a less significant way in terms of numbers. It is important to note that the distribution owes its origin in part to the fact that vapor diffusion methods have been available to the community for the longest period. In contrast, the bicelle method is a relative newcomer.

3.3.3 Detergent Used in Solubilization/NMR

A total of 43 different detergents have been used in protein structure determination studies across all of the methods represented in the MPDB. Octyl glucoside is by far the most extensively used, with LDAO taking up a close second position. C_8E_4, dodecyl maltoside, and decyl maltoside are in a lesser used group followed by $C_{12}E_9$ and dodecyl phosphatidylcholine. The distribution changes little when the statistical analysis is limited to XRD-based structures (table 1). However, for structures solved by NMR the dominant detergents include dodecyl phosphatidylcholine, sodium dodecyl sulfate and dihexanoyl phosphatidylcholine. These facilitate solution structure determination as mixed micelles.

3.3.4 Experimental Technique

The default statistical analysis shows the distribution of MPDB records across the four structure techniques. XRD accounts for 77%, while NMR represents 20%. The remaining records are shared between cryo-EM (1%) and electron diffraction (2%). Other search combinations under **Experimental Technique** are possible but generally yield just single values.

3.3.5 Expression System

Across all experimental techniques, *Escherichia coli* stands out as the most frequently used expression system with which to produce recombinant membrane proteins for structure determination (table 1). *Rhodobacter sphaeroides* and *Halobacterium salinarum* are also very popular, reflecting the fact that expression systems have been in place for both for some time. Less popular but still with representation within the body of structures are the eukaryotes, *Pichia* (yeast) and *Spodoptera* (insect cells). When examined from the point of view of XRD structures, the profile is essentially the same. The most important ‘expression system’ providing structures by the NMR route is ‘synthetic’ as man-made peptides.

3.3.6 Journal

Over the entire database, the major journals publishing structures of membrane protein and peptides include *J. Mol. Biol. (JMB)* followed successively by *Science*, *Nature* and *Biochemistry*. When the analysis is performed on XRD-based records, the order changes to *Science, JMB* and then *Nature* with *Biochemistry, Structure, PNAS,* and *J. Biol. Chem.* taking up the slack. When applied to NMR structures, the most popular outlets are Biochemistry followed by *JMB* and *Eur. J. Biochem.,* in that order.

3.3.7 Lipid Used in Bilayer Crystallization/NMR

This refers to the bicelle, cubic phase and vesicle fusion methods for growing crystals. The most commonly used lipid within the group is monoolein, the lipid upon which the cubic phase method was founded. Dimyristoyl phosphatidylcholine is next on the list and is used exclusively for bicelle crystallogenesis. The third, most commonly used lipid in this category is referred to as the purple membrane lipids. They were employed to grow bacteriorhodopsin crystals by the vesicle fusion method.

3.3.8 pH

When performed over the entire database, the pH range used in successful membrane protein and peptide structure determinations extends from pH 1.4 to 10 (fig. 6). The most frequently used pH value (the mode) is 7.0, with large numbers of successes registered at pH 5.6, 6.0, 6.5, 7.5 and 8.0. Since structures determined by

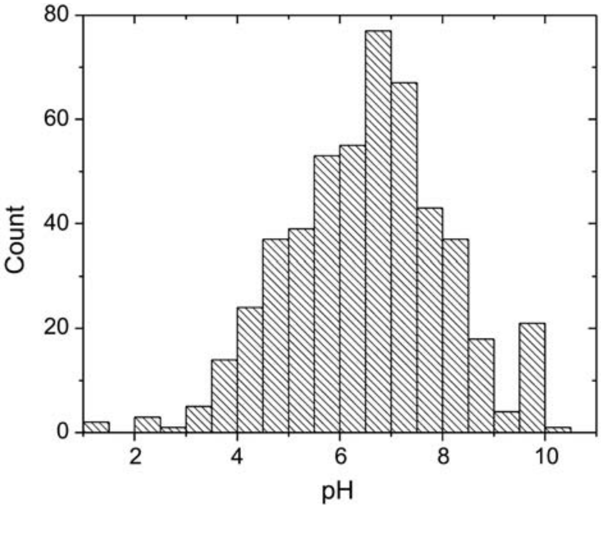


Figure 6. pH dependence of the number of records in the MPDB across all structure determination methods.

XRD represent ~80% of the records in the MPDB, the distribution is heavily weighted by the characteristics of this subset. Thus, repeating the analysis but limiting it to XRD-based structures gives a distribution similar to that obtained for the entire database.

The non-binned distribution that appears on the website under **Statistics**, with spikes at regular intervals of pH, very likely reflects the preferred (over-) sampling of the pH scale by investigators and by the use of crystallization kits with set pH values. The question arises then as to the utility of the distribution in informing crystallogenesis trials. It does highlight the fact that the crystallization of membrane proteins and peptides occurs over a very wide range of pH values. Thus, unless the protein loses activity, denatures or undergoes chemical modification, or the lipids, additives and precipitants used for protein solubilization and/or crystal growing are unstable, then it is appropriate to investigate the full range of pH values.

3.3.9 Precipitant Used in Crystallization

The panel of precipitants used in crystallization has 44 entries. The most commonly employed is PEG 4000, with PEG 400 a distant second (table 1). Ammonium sulfate and PEG 2000 vie for third place. Sodium potassium phosphate is next on the list, followed by potassium phosphate and PEG 600. This is across all crystallization methods that are dominated by the vapor diffusion methods. When the statistics are run on the cubic phase method, sodium potassium phosphate is by far the most popular precipitant. Potassium and sodium chloride and PEG 4000 have also been used but with considerably lower frequency. The pattern is quite similar when performed on the batch crystallization method.

3.3.10 Source Organism

The most frequently used source organism for all structures in the MPDB is *Escherichia coli*. This reflects the popularity of the organism as a source of native membrane proteins as well as a system for homologous expression. The next most popular in order of frequency are *Halobacterium salinarum* and *Rhodobacter sphaeroides* (table 1). These, in turn, are sources for the bacterial rhodopsins and the photosynthetic reaction centers, and their high count makes sense given the intensity with which these respective systems have been investigated. The pattern changes little when the statistics are run on XRD structures alone. However, when looked at from the NMR-based structures perspectives, the human source emerges with highest frequency.

3.3.11 Temperature

The temperatures used to solve protein and peptide structures across all methods in the MPDB range from a low of 4 to a high of 60 °C. The mode is 20 °C, with 4 °C assuming second place. The next most commonly used temperatures lie above and below the mode with a distribution that is skewed to the high temperature side. The distribution for XRD-based structures, as expected, is similar to that described, but the highest temperature employed drops to 37 °C (table 1). The mode in the case of NMR-based structures is 30 °C, with significant numbers of data being collected above 20 °C and up to 55 °C. This reflects the fact that the bulk of the NMR structures refer to peptides rather than proteins and that spectral quality generally improves with temperature.

3.3.12 Year

The cumulative number of protein and peptide structures in the MPDB, across all experimental techniques for structure determination, has grown in a manner that is close to exponential in the period 1982 through 2004 (fig. 7). (Note that what is represented here is the Publication Year for the primary citation. As commented on above, it can differ significantly from the Deposition Date.) A similar function can be used to describe growth in the number of XRD-based structures. 2001 was a bumper year for NMR structures, and the following 2 years have witnessed progressively smaller numbers of entries. However, based on the numbers thus far, 2005 looks like it will witness a very sizable output of such structures. It is important to note that the vast majority (88%) of NMR

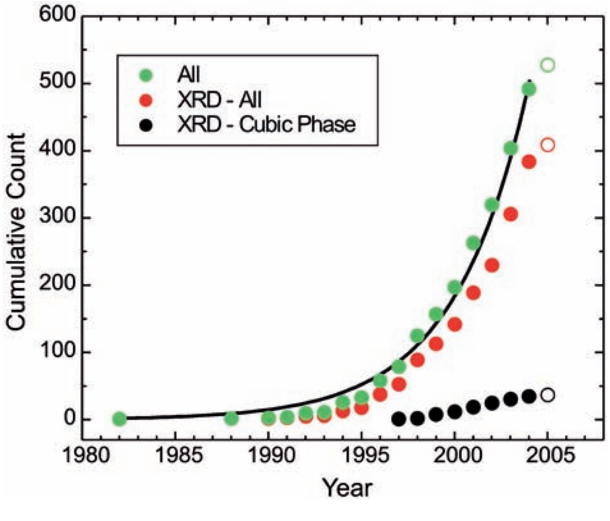


Figure 7. Dependence of the cumulative MPDB record count on the year of primary citation publication. The analysis was performed on all records in the MPDB (blue-green), on XRD records across all crystallization methods (red), and on XRD records in the lipidic cubic phase method subset (black). An exponential best fit to data in blue-green in the period 1982–2004 is shown as a solid black line. The equation for the line takes the form Y = A $e^{-B\,(X-1982)}$ where Y is cumulative count, X is year, A is 2.0 and B is 0.251. Note that the data for 2005 only include records for the first 6 months of that year and are identified by open symbols.

records refer to peptides. The authors have a particular interest in the lipidic cubic phase method for crystallizing membrane proteins. The corresponding annual statistics and those for all XRD-based structures are also shown alongside the overall membrane protein/peptide structure statistics in figure 7.

3.4 Updating

The goal is to update the MPDB on a weekly basis. This will take place in parallel with the PDB, which is also a weekly event. A program has been written to query all PDB records that have been entered in the past week for those that contain the word 'membrane'. Those records deemed appropriate by MPDB staff are processed, and relevant data and links are entered into the MPDB using a web interface. The database is time-stamped with the latest date on which updating occurred, and all statistics and current holding values are recalculated and posted. At the same time, PDB IDs that have been superseded are replaced in the MPDB with the new ID code. New records are compared with those carrying the 'To Be Published' flag and appropriately annotated.

3.5 Recommendations

First and foremost, the MPDB is a convenient and efficient resource of information on membrane proteins and peptides. The bulk of the entries in the database refer to integral membrane proteins whose structure was solved by XRD. The latter subset of records can be mined for conditions that are compatible with successful crystallization. What follows is a summary of the conditions that have been most useful in this regard across all such entries in the data bank. These include pH: 4.5–8.5; additives: sodium chloride, heptanetriol, glycerol, MPD, magnesium chloride; detergent: octyl glucoside, LDAO, C_8E_4, dodecyl maltoside, decyl maltoside, $C_{12}E_9$; precipitant: PEG 4000, PEG 400, PEG 2000, ammonium sulfate, potassium phosphate, sodium potassium phosphate; temperature: 4 °C, 18–25 °C.

It is appreciated that the above set of conditions has been culled from across all XRD records in the MPDB without regard to source organism, membrane protein type and so on. However, given an entirely new protein with no known homologs in the MPDB, the above conditions represent a good starting point with which to begin the search for structure-grade crystals.

3.6 The Future

The MPDB in its current incarnation represents a compromise of sorts. There is always a lot more that can be done to improve its appearance, functionality and scope. But each takes time and effort. Our hope is that what is being released in parallel with this publication is useful as it stands. We realize that there are errors and omissions, although we have made every effort to keep these to a minimum. Users are requested to bring these to our attention for correction and to submit recommendation for how to improve the resource.

In the interests of time, several protein characteristics are not in the current version of the MPDB. These include the organelle (mitochondrion, chloroplast, lysosome, nucleus, etc.) and membrane (endoplasmic reticulum, bacterial outer membrane, plasmamembrane, sarcoplasmic reticulum, rod outer segment, 'inclusion body', etc.) source. The intent is to have this information in the next version of the data bank.

Currently, MPDB records house data on the major chemical components used to solve the structure of a given protein. The plan for the future is to augment this identity information with actual concentrations of the assorted materials used to effect crystallization, solubilization and dispersion. This will better lend itself to data mining for precise conditions supporting crystallization of particular membrane protein types.

The statistics feature in the current MPDB is useful but is limited by the fixed number of searches the curators have made available. A later version of the data bank will provide the user with the ability to perform personalized statistical analysis on the data and to create online their own graphical representation of the results of a search that can be ported into a text document.

4 Conclusions

A web-based database of membrane protein and peptide structure and function, the MPDB, has been released. It is updated weekly in parallel with the PDB. The database is searchable based on a host of criteria ranging from protein name and source organism to the method used to crystallize the protein for use in diffraction measurements. The vast majority of records in the database refer to integral membrane proteins whose structure was solved by crystallographic means. Statistical analysis can be performed on the data online, and this feature has been used to identify conditions suitable for beginning crystallization trials on new membrane proteins.

*Acknowledgements***.** The authors thank O. Slattery for assistance preparing an early version of the data bank. J. Clogston, J. Lyons, N. Hoefer, O. Slattery, W. Liu and Y. Misquitta are gratefully acknowledged for testing the data bank and for comments on the manuscript. We are indebted to K. Kadyshevskaya for creating the MPDB logo. Supported in part by Science Foundation Ireland (02-IN1-B266), the National Institutes of Health (GM61070 and GM075915), and the National Science Foundation (DIR9016683 and DBI9981990).

1 Berman H. M., Westbrook J., Feng Z., Gilliland G., Bhat T. N., Weissig H. et al. (2000) The Protein Data Bank. Nucleic Acids Res. **28:** 235–242

2 Tusnády G. E., Dosztányi Z. and Simon I. (2004) Transmembrane proteins in the Protein Data Bank: identification and classification. Bioinformatics **20:** 2964–2972

3 Caffrey M. (2003) Membrane protein crystallization. J. Struct. Biol. **142:** 108–132

4 Michel H. (1982) Three-dimensional crystals of a membrane protein complex. The photosynthetic reaction centre from Rhodopseudomonas viridis. J. Mol. Biol. **158:** 567–572

5 Diesenhofer J., Epp O., Miki K., Huber R. and Michel H. (1985) Structure of the protein subunits in the photosynthetic reaction center of Rhodopseudomonas viridis at 3 Å resolution. Nature **318:** 618–624

6 Deisenhofer J., Epp O., Sinning I. and Michel H. (1995) Crystallographic refinement at 2.3 Å resolution and refined model of the photosynthetic reaction centre from Rhodopseudomonas viridis. J. Mol. Biol. **246:** 429–457

To access this journal online:
http://www.birkhauser.ch