Research Paper

Biological Activity Database

Kazuo Satake, Yu Watanabe, and Akira Tsugita
Japan International Protein Information Database (JIPID) Research Institute for Biosciences, Science University of Tokyo

As a complement to the protein sequence database of PIR and the tertiary structural data bank of Brookhaven, we will present protein biological activity databases. The BAD's contain information regarding the functional properties of more than 5000 kinds of enzymes. The binding properties of the carrier, inhibitor, or modulator proteins are also included in these databases. In addition, information about wild-type proteins will be presented together with information concerning variant molecules that may be compared under the same analyzing conditions.

1. Introduction

The support given to recent initiatives to elucidate the complete genomic sequences of organisms, such as yeast, rice, and human, demonstrate that these data are generally recognized to have significant scientific and application value. Concurrent with the growing requirement for biological macromolecule sequence data is the demand for associated information concerning the biological properties of these molecules. These data are fundamental for comprehending the biological information encoded in biomacromolecuele structures. Further, progress in the basic biosciences suggest almost immediate biotechnological applications such as diagnosis and cure of genetic diseases, the design of novel enzymes and drugs, and the introduction or enhancement of desirable characteristics in livestock and crops.

Sequence databases for both nucleic acid and proteins and protein tertiary structure database play an integral part bioscience research by acting as repositories for elucidated structure, providing software tools for the evaluation of data, and facilitating the dissemination of these data to the scientific community. With the aim of providing a more extensive data library to complement the sequence data, PIR-International, and tertiary structure data, PDB, is introducing several novel databases. The one of these databases, a Biological Activity Database (BAD) contains information pertaining to the biological activity of proteins. Initially the information will be restricted to those proteins already having amino acid sequence entries in PIR, and the entries in PDB.

The functional protein may contain several subunits or domains, each with a separate entry in PIR. Also PIR entry may be of the precursor protein. The functional proteins in BAD is essentially the same as those in PDB but not always. Consequently there is not a direct one-to-one correlation between entries in PIR or PDB and BAD.

An entry in these non-sequence databases basically follow the recommendations made in the CODATA standardized data exchange format[1], in which information is divided up into distinct data items. Each data item may be accessed independently from the file it

Table 1. Data items in the Sequence Database, Biological Activity Database(BAD)

PIR	Biological Activity	
	General file	Specific file
HEADER	HEADER	HEADER
ENTRY-CODE	ACCESSION	ACCESSION
TITLE	TITLE	TITLE
EC-NUMBER	EC-NUMBER	EC-NUMBER
	SYSTEMATIC-NAME	SYSTEMATIC-NAME
ALTERNATE-NAME	ALTERNATE-NAME-G	ALTERNATE-NAME
INCLUDES		
	DATE	DATE
SOURCE		SOURCE
	DISTRIBUTION	
HOST		HOST
COMMENT		
	ORGANIZATION-G	ORGANIZATION
	ACTIVE-RESIDUES	
	BINDING-RESIDUES	
	REACTION	REACTION
	ASSAY	
	FUNCTION FUNCTION	
		POST-TRANSLATIONAL
	APPLICATION-G	APPLICATION
	COMMENT-G	
GENETIC		GENETIC
KEYWORDS		
FEATURE		FEATURE
SUMMARY		UMMARY
SEQUENCE-P		SEQUENCE-P
CROSS-REFERENCE	CROSS-REFERENCE(S)	CROSS-REFERENCE

is contained in. This allows the databases to be organized into a network structure or data-base web, whereby data items in any one database may be transferred to any other database in the web. Consequently, overlap of data between the databases will be kept to a minimum. The initial growth of these databases will be controlled by the introduction of specific parts of the database at different times.

Software for the manipulation of these databases is written in C language to maximize the portability of the database. As the database is written in plain text format, it may be easily converted to HTML (Hyper Text Markup Language), standard format of the WWW (World Wide Web). The database will be accessible on WWW in the future. The BAD has been initially constructed to compliment to PIR[2], and extended to PDB, as published in[3].

2. The Biological Activity Database(BAD)

The BAD currently contains information on enzymes, electron-carrier proteins, and oxygen-carrier proteins. Other proteins will be subsequently entered. The database has two levels of organization; General Datafiles and Specific Datafiles. Common information about proteins are summarized and maintained in the General Datafiles, thus avoiding repetition in the Specific Datafiles. The Specific Datafiles contain information about proteins from particular biological sources. Table 1 shows the data items in PIR, the BAD for both General and Specific entries. Data items in bold are entered into the database directly, while unbold data items may be transferred from another entry either with or without minor modifications. For example, the TITLE data item in the Specific Datafile may be transferred from the PIR entry.

2.1 The General Datafile

Each entry in the general Datafile may refer to one or more entries in the Specific Datafile. Descriptions of the data items used in the General Datafile format are described below:

The HEADER 'KG' denotes that the entry is a General Datafile for enzymes, while 'KH' denotes for the others than enzymes.
An ACCESSION is assigned to each entry in the database. It acts as a tag, is entry specific, and contains no informational content.
The TITLE data item is the name of the protein. The name, where feasible, takes the title of the corresponding PIR or PDB entry TITLE. In the case of enzymes the TITLE is the same as the trivial name in Enzyme Nomenclature which is also usually used in PIR entry.

The EC-NUMBER data item is taken directory from Enzyme Nomenclature. The SYSTEMATIC-NAME data item is used for enzymes or others, and is taken from Enzyme Nomenclature or other standardized nomenclatures if exist.
The ALTERNATE-NAME data item contains the altarnate names of the protein found in PIR, PDB, Enzyme Nomenclature and the literatures.
The DISTRIBUTION data item denotes the spatial and temporal distribution of the protein.
The ORGANIZATION-G data item records any common structural organization of the protein in biological active form.
The COFACTORS-G data item names any general cofactors necessary for the function of the protein such as prosthetic group, coenzyme, metal ion or haem group.
The ACTIVE-RESIDUES data item denotes any residues or motifs that commonly act as residues or sequences.
The BINDING-RESIDUES data item denotes any residues that commonly act as binding residues for phosphates, carbohydrates, nucleic acids, metals, etc.
The REACTION data item describes the reaction involving the protein. For enzymes the reaction may be partly taken from Enzyme Nomenclature.
The ASSAY data item describes the recommended method to measure activity.

The FUNCTION data item contains information related to the function of the molecule. This item includes the subidentifiers; # SUBSTRATE which catalogues common substrates (listed in Enzyme Nomenclature) used by an enzyme, # PH-RANGE which denotes the range of pH, functionally active, # ACTIVATOR-G and # INHIBITOR-G which denote molecules that activate or inhibit the function, and # ACTIVITY-G which contains free text comments of biological activity.

The APPLICATION-G data item contains comments about the industrial or clinical applications.
The COMMENT-G data item contains any miscellaneous information not covered by the other data items.
The CROSS-REFERENCES-G refers to the entries in the Specific Datafiles only. The end-of-entry data item consists of three consecutive slashes '///'.

The inclusion of General Datafile entries in BAD serves several purposes. It allows

Table2. Function data items of Specific Datafiles in Biological Activity Database

Enzyme	Enzyme Modulater	COAGULATION FACTOR	COMPLEMENT	Electron -Carrier	Oxygen -Carrier	Hormon -Carrier
SUBSTRATE						
INHIBITOR	INHIBITOR					
ACTIVATOR	ACTIVATOR	ACTIVATOR	ACTIVATION			
EQUILIBRIUM -CONSTANT		(ACTIVATION)				
ACTIVITY		ACTIVITY	ACTIVITY			
OPTIMAL-PH			#RECEPIOHTION			
CONDITIONS			#DEGRADATION			
	PHYSIOLOGY	PHYSIOLOGY	PHYSIOLOGY			PHYSIOLOGY
		PATHOLOGY	PATHOLOGY			PATHOLOGY
				REDOX POTENTIAL		
				CONTDITIONS		
					OXYGEN AFFINITY	
					BOHR EFFECT	
					COOPERANTIVITY	
					ANION EFFECT	
					CATION EFFECT	
					CARBON DIOXID EFFECT	
					HEAT OF OXYGENATION	
					CONDITIONS	
						LIGAND SPECTRUM
						BINDING CONSTANT
						CONDITIONS

common information about proteins to be maintained separately from biological source-specific information thereby reducing overlap of information in the database. In addition to common information, the General Datafile contains a directory of all the biological sources of the proteins as well as their Specific Datafile codes, while the General Datafile has not always been made corresponding to the respective Specific Datafile. In the case of enzymes, the General Datafiles act as an on-line enzyme handbook. They contain lists of substrates used by the enzymes, as well as the information presented in Enzyme Nomenclature.

2.2 The Specific Datafile

Specific Datafiles describe the biological activities of functional mature proteins from particular biological sources. Information in the data items from the General Datafiles and other databases in the web may be transferred to Specific Datafiles:

The HEADER 'KS' denotes that the entry is a Specific Datafile entry for enzymes, while 'KT' denotes for the others than enzymes. An ACCSSION is assigned to each entry in the database. It contains no informational content.

The TITLE data item is essentially taken from PIR or PDB entry. This data item includes the name of the biological source and in the case of enzymes, the EC number. The data item is altered if the sequence title is that of the precursor molecule or subunit.

Table 3. General Datafile

ACCESSION-NUMBER	KJ0322
TITLE	DNA-directed RNA polymerase
EC-NUMBER	2.7.7.6
# Clasification	This enzyme belongs to the transferases that transfer phosphorous-containing groups -a diphosphotransferase
SYSTEMATIC-NAME	DNA Nucleoside-triphosphate: RNA nucleotidyltransferase (DNA-directed)
ALTERNATE-NAME-G	RNA nuceotidyltransferase (DNA-directed), Transcriptase
DATE	27-Sep-1989
DISTRIBUTION	Nucleus, chloroplast
ORGANIZATION-G	2 alpha, 1 beta, 1beta', 1 sigma
# Description	The sigma subunit dissociates from the core enzyme 2 alpha, beta, beta'
REACTION	n Nucleoside triphosphte = n pyrophosphate + RNA n
# Comment	The reaction is reversible.
ASSAY	
SUBSTRATE	ATP, CTP, GTP, UTP, Mn2+, Mg2+, Pi
PH-RANGE	7.5-7.6
INHIBITOR	Polynucleotides (in rat testis)
COMMENT-G	The enzyme needs DNA as a template.
CROSS-REFERENCE(S)	RNBP17 Bacteriophage T7, RNBPT3 Bacteriophage T3, RNVZ22 Vaccinia virus (strain WR), RNECA Eschrichia coi, RNNTA Common tobacco chloroplast, RNLVA Liverwort (Marchantia polymorpha) chloroplast, RNECB Escherichia coi, RNECB2 Escherichia coli, RNNTB Common tobacco chloropast, RNLVB Liverwort (Marchantia poymorpha) chloroplast RNBY2L Yeast (Saccharomyces cerevisiae), RNFF2L Fruit fly, RNBY3L Yeast (Saccharomyces cerevisiae), RNVZ47 Vaccinia virus (strain WR), RNECC Escherichia coli, RNECC2 Escherichia coli, RNECS Escherichia coli, RNEBST Salmonella typhimurium, RNECO Escherichia coli
///	

The SYSTEMATIC-NAME is taken directly from the General Datafile.

The ALTERNATE-NAME is taken from PIR entry but additional names may be added. In some cases, even naming of antigen or factor are added.

The SOURCE data item denotes the scientific and common names of the biological source of the protein as well as designations for strain, plasmid, clone, tissue isolated from, life cycle expressed, description of source, and taxonomic classification. This information may be taken from PIR or PDB entry.

The HOST data item is taken directly from the General Datafile.

The ORGANIZATION data item describes the structural organization of the biological active molecule.

The REACTION data item is taken directly from the General Datafile.

The FUNCTION data item denotes data relating to the function. The subidentifiers for this data item in a few biological fields are listed in Table 2.

The POST-TRANSLATIONAL data item denotes any post-translational modifications.

The APPLICATION data item describes industrial or clinical applications. When appropriate, this data item is transferred from the General Datafie.

The GENETIC data item contains genetic information associated with the entry. This data item is transferred from PIR.

The FEATURE data item denotes the sites or regions of biological interest and is taken from PIR or PDB. The descriptors used for this data item include: active site, binding site, disulfide bond, and selected items for domain, duplication, modified site, and region. These data may be transferred from the feature table in PIR. While residue number specifications in PIR are often for precursor molecules, the residue numbers for the mature proteins in BAD are needed to change. Thus they are automatically adjusted from PIR entry using information in the ORGANIZATION data item. Information in the ACTIVE- and BINDING-RESIDUES data items used in the General Data file are also replaced by the FEATURE data items.

The SUMMARY data item contains information about the molecular weight of the mature protein and its amino acid composition automatically generated from PIR into the mature protein.

The SEQUENCE-P data item displays the amino acid sequence(s) of the biologically functional mature protein. This data is automatically generated from PIR, using information in the ORGANIZATION data item. When a protein is composed of two or more peptide chains, all sequences will be displayed.

The CROSS-REFERENCE data item denotes cross-references to corresponding entries in other databases that are not in the database web. This data item is partly taken from PIR entry.

The end-of-entry data item marks the end of the entry by the identifier consisting of '///'.

Examples of the formats for the General and Specific Datafiles are given in Table 3 and 4. BAD describes biologically active mature proteins, and do not always show one-to-one correlations with entries in PIR and /or PDB. This is because the biologically active mature proteins described in the non-sequence databases may range from simple monomers to more complex heteromers. Also the polypeptide chain may contain several enzymic functions.

The ORGANIZATION data item aims to describe the structural organization of the mature protein. It consists of the number of each subunit, the name of the subunits and residue numbers of the mature polypeptide followed by the sequence entry code in square brackets, i.e., subunit number subunit name ;residues[PIR code] .

Escherichia coli DNA-directed RNA polymerase, consists of a core enzyme and a sigma chain. The core enzyme is composed of two alpha chains, one beta chain, and one beta' chain. The ORGANIZATION data item is written as:

2 alpha;1-329[RNECA]
1 beta;1-1342[RNECB]
1 beta';1-1407[RNECC]
1 sigma;1-613[RNECS]

In some cases, a polypeptide chain in the PIR database consists of a precursor or biologically inactive form of the molecule. In BAD, the ORGANIZATION data item only catalogues the polypeptide after processing, although the archive and code to the original PIR entry is maintained. The display of the sequence of the mature protein when it differs from the precursor form is renumbered. For example the trypsinogen entry for dog, TRDGC, consists of 246 residues, however the trypsin enzyme obtained from the same polypeptide contains only 223 residues.

1;7-131[TROBTR]
1;132-229[TROBTR]

The Specific Datafiles describe the organization, and biological activity of functional mature proteins from particular biological sources. The Specific Datafiles contain information about wild-type molecules. These data may be used for comparison with data about variant molecules that are stored in the Artificial Variant Data-base (VAD)[4]. Mature sequences of the protein can be generated from the original sequence database entries.

References

[1] Gorge,D.G., Mewes,H.W., and Kihara H. Protein Seq Data Anal. Vol.1, pp. 27-39(1987)

[2] Jone,. C.S., Tsugita, A., Satake, K., Okibayashi, F., Imai, K., Yagi, T., Takahashi, K., and Yeh L.S. Protein Seq Data Anal. Vol.4, pp.367-374(1991)

[3] Satake, K., Miyazaki, K., Ubasawa, A., Shen, R., and Tsugita, A. Extended Abstracts and Proceedings 15th Intnal, CODATA Confer. (ed. by Glaser, P.S., Prado, C., and Tsugita, A) pp.188(1996)

[4] Ubasawa, A., Okibayashi, F., Jone, C.S., Ikehara, M., Gorge, D.G., and Tsugita,A., Protein Seq Data Anal Vol.4, pp.341-347(1991)

(Received January 9, 1998)

著者紹介

佐竹 一夫
理学博士
東京理科大学 JIPID 顧問
元東京理科大学理学部教授
日本化学学会, 薬学会, 生化学会
有機合成化学協会等会員

渡辺 祐
東京理科大学大学院生命科学研究科
修士課程 在学中

次田 晧 (正会員)
理学博士
東京理科大学教授
元大阪大学医学部教授
元 CODATA 副会長、元日本蛋白工学会会長
生化学会, 生物物理学会, 分子生物分会等会員
EMBO(Europe13ヶ国立分子生物研究所) 正会員
スイスバーゼルアカデミー会員

Table 4. Specific Datafile

```
ACCESSION-NUMBER      KJ0011
TITLE                 DNA-directed RNA polymerase (EC 2.7.7.6) - Escherichia coli
SYSTEMATIC-NAME       DNA Nucleoside-triphosphate: RNA nucleotidyltransferase
                      (DNA-directed)
ALTERNATE-NAME        RNA nuceotidyltransferase (DNA-directed), Transcriptase
SOURCE                Escherichia coli
DATE                  27-Sep-1989
LITERATURE            L01031
AUTHORS               Ovchinnikov, Y.A., Lipkin, V.M., Modyanov, N.N., Chertov, O.Y.,
                      and Smirnov, Y.V.
JOURNAL               FEBS Lett. 76, 100-111, 1977 <OVC>
LITERATURE            L01032
AUTHORS               Gentry, G.R., and Burgess, R.R.
JOURNAL               Gene 48, 33-40, 1986 <GEN>
LITERATURE            L01001
AUTHORS               Barman, T.E.
JOURNAL               Enzyme Handbook, 1, 467, 1969, <BAR>
ORGANIZATION          2 alpha; 1-329[RNECA],1 beta;1-1342[RNECB],
                      1 beta';1-1406[RNECC], 1 sigma;1-613[RNECS],
                      1 omega;1-91[RNECO]
                      The core enzyme consists of 2 alpha, 1 beta and 1 beta'
                      subunits. The holoenzyme has an additional sigma subunit.
                      <OVC> The omega subunit is not required for transcription
                      but copurifies with RNA poymerase. <GEN>
REACTION              n nucleoside triphosphate = n pyrophosphate + RNA n
                      The reaction is reversible.
FUNCTION
  #pH                 7.5(Tris)
  #Temperature        38C
  #Others             calf thymus DNA primer
Substrate             Km(M)
ATP                   8xE-6   <BAR>
CTP                   9xE-6   <BAR>
GTP                   6xE-6   <BAR>
UTP                   6xE-6   <BAR>
Mn2+                  2xE-6   <BAR>
Mg2+                  4.6xE-6 <BAR>
Pi                    1xE-6   <BAR>

SEQUENCE

PIR1:RNECA
DNA-directed RNA polymerase (EC 2.7.7.6) alpha chain - Escherichia coli

              5         10        15        20        25        30
  1 M Q G S V T E F L K P R L V D I E Q V S S T H A K V T L E P
 31 L E R G F G H T L G N A L R R I L L S S M P G C A V T E V E

271 K A E A I H Y I G D L V Q R T E V E L L K T P N L G K K S L
301 T E I K D V L A S R G L S L G M R L E N W P P A S I A D E

PIR1:RNECB
DNA-directed RNA polymerase (EC 2.7.7.6) beta chain - Escherichia coli

              5         10        15        20        25        30
  1 M V Y S Y T E K K R I R K D F G K R P Q V L D V P Y L L S I
 31 Q L D S F Q K F I E Q D P E G Q Y G L E A A F R S V F P I Q
```

```
1291 L T V K S D D V N G R T K M Y K N I V D G N H Q M E P G M P
1321 E S F N V L L K E I R S L G I N I E L E D E

PIR1:RNECC
DNA-directed RNA polymerase (EC 2.7.7.6) beta' chain (version 1) -
    Escherichia coli

              5         10        15        20        25        30
   1 M K D L L K F L K A Q T K T E E F D A I K I A L A S P D M I
  31 R S W S F G E V K K P E T I N Y R T F K P E R D G L F C A R

1351 V I V G R L I P A G T G Y A Y H Q D R M R R R A A G E A P A
1381 A P Q V T A E D A S A S L A E L L N A G L G G S D N E

PIR1:RNECS
transcription initiation factor sigma 70 - Escherichia coli

              5         10        15        20        25        30
   1 M E Q N P Q S Q L K L L V T R G K E Q G Y L T Y A E V N D H
  31 L P E D I V D S D Q I E D I I Q M I N D M G I Q V M E E A P

 571 Y T L E E V G K Q F D V T R E R I R Q I E A K A L R K L R H
 601 P S R S E V L R S F L D D

PIR1:RNECO
DNA-directed RNA polymerase (EC 2.7.7.6) omega chain - Escherichia coli
              5         10        15        20        25        30
   1 M A R V T V Q D A V E K I G N R F D L V L V A A R R A R Q M
  31 Q V G G K D P L V P E E N D K T T V I A L R E I E E G L I N
  61 N Q I L D V R E R Q E Q Q E Q E A A E L Q A V T A I A E G R
  91 R
```