ProTherm and ProNIT: thermodynamic databases for proteins and protein–nucleic acid interactions

M. D. Shaji Kumar[1], K. Abdulla Bava[1], M. Michael Gromiha[2], Ponraj Prabakaran[3], Koji Kitajima[1], Hatsuho Uedaira[4] and Akinori Sarai*

Department of Bioscience and Bioinformatics, Kyushu Institute of Technology (KIT) 680-4 Kawazu, Iizuka, 820-8502 Japan, [1]Advanced Technology Institute, Inc. (ATI), 2-3-13-103 Tate, Shiki, Saitama 353-0006, Japan, [2]Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), AIST Tokyo Waterfront Bio-IT Research Building, 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan, [3]Laboratory of Experimental and Computational Biology, NCI, NIH, Frederick, MD 21702, USA and [4]Tsukuba Materials Information Laboratory, 3-23-4, Ninomiya, Tsukuba, 305-0051 Japan

ABSTRACT

ProTherm and ProNIT are two thermodynamic databases that contain experimentally determined thermodynamic parameters of protein stability and protein–nucleic acid interactions, respectively. The current versions of both the databases have considerably increased the total number of entries and enhanced search interface with added new fields, improved search, display and sorting options. As on September 2005, ProTherm release 5.0 contains 17 113 entries from 771 proteins, retrieved from 1497 scientific articles (∼20% increase in data from the previous version). ProNIT release 2.0 contains 4900 entries from 273 research articles, representing 158 proteins. Both databases can be queried using WWW interfaces. Both quick search and advanced search are provided on this web page to facilitate easy retrieval and display of the data from these databases. ProTherm is freely available online at http://gibk26.bse.kyutech.ac.jp/jouhou/Protherm/ protherm.html and ProNIT at http://gibk26.bse. kyutech.ac.jp/jouhou/pronit/pronit.html.

INTRODUCTION

Thermodynamic database for proteins and mutants (ProTherm) and thermodynamic database for protein–nucleic acid interactions (ProNIT) are two comprehensive, integrated databases that document experimentally determined thermodynamic parameters published in the literature. Both ProTherm (1–4) and ProNIT (5) include several thermodynamic parameters along with sequence and structural information, experimental methods and conditions, and literature information. Recent years have seen tremendous progress in studies on proteins owing to the development of various experimental methods to analyze proteins at the genome scale. The correlation between structure and thermodynamics of these key molecules provides valuable insights into the way in which they function. Even though the information is available in scientific journals, books (6,7) and literature databases, retrieving useful, specific data from these resources is time consuming and laborious. Our major goal in developing these databases is to provide the scientific community a single, comprehensive data repository, where all the thermodynamic data related to protein stability and protein–nucleic acid interactions are available. The availability of such thermodynamic databases would be a valuable resource for understanding the protein folding mechanism, protein stability, molecular recognition and gene expressions. This can lead to a wide spectrum of applications such as developing algorithms/ methods for prediction systems, protein engineering and quantitative simulation of gene regulatory networks. The thermodynamic data available in ProTherm and ProNIT are widely used by researchers to study the underlying mechanisms of protein stability upon mutations and protein–nucleic acid interactions (see the reference sections on both the websites). This paper describes the major updates and enhancements to these databases for the last few years.

CONTENT, ORGANIZATION AND DATA COLLECTION

Both the databases contain information on protein, mutational information, experimental methods and conditions, several thermodynamic parameters and literature information.

Table 1. Contents of ProTherm and ProNIT

Contents of the databases ProTherm	ProNIT
Protein information	Protein information
Name, Source	Name, Synonyms
PIR, SWISSPROT	Source, Sequence
PDB code	EC, PIR, SWISSPROT
EC, PMD number	PDB code
Mutation details	Biological unit
Secondary structure	Mutation details
Accessible surface area (ASA)	Secondary structure
Experimental condition:	ASA
Temperature	Nucleic acid information:
pH	Name
Buffer, Ion	Source
Protein concentration	Type (DNA or RNA)
Measure (DSC, CD and so on)	Sequence (wild and mutant)
Method of denaturation	Mutation details
Thermodynamic data:	GenBank Number
Denaturant denaturation:	Complex information:
Free energy of unfolding: ΔG_{H_2O}	PDB code, NDB code
Difference in ΔG_{H_2O} : $\Delta\Delta G_{H_2O}$	Conformation of protein
Denaturation concentration: C_m	Conformation of Nucleic Acid
Slope of denaturation curve: m	ASA
Temperature: T	Experimental condition:
Thermal denaturation:	T, pH, Buffer, Ion, Additives
Free energy of unfolding: ΔG	Experimental method
Difference in ΔG: $\Delta\Delta G$	Binding data:
Transition temperature: T_m	Dissociation constant: K_d
Change in T_m: ΔT_m	Association constant: K_a
Enthalpy change: ΔH_{cal}, ΔH_{vH}	Free energy change: ΔG
Heat capacity change: ΔC_p	Enthalpy change: ΔH
Literature:	Heat capacity change: ΔC_p
Reference, Author	Literature:
Keywords, Remarks	Reference, Author
Related entries	Keywords, Remarks
	Related entries

Previous publications (1–5) explain in detail the content and organization of the databases. Table 1 summarizes the contents of ProTherm and ProNIT. ProTherm and ProNIT are implemented in 3DinSight (8), a relational database system for structure, function and property of biomolecules. This facilitates more efficient search and retrieval of data by flexible queries, and enables users to gain insight into the relationship among structure, thermodynamics and function of proteins. We have been collecting the thermodynamic data from published original articles, by searching the PubMed literature database with a combination of specific terms, as well as by searching online journals probably containing thermodynamic data. The database does not contain any predicted or computational interaction data. Researchers then extract the relevant data from the selected articles. The input data are checked automatically by checking programs and also manually to avoid errors. Then, we upload the data first to a test site, where expert curators check and verify the data. After this checking, we upload the data to the public site for users. Furthermore, an email notification for each input entry is sent to the corresponding author, which enables the authors to check their own data and thereby improve the data validation.

DATABASE STATISTICS

We update both the databases frequently. The current release, ProTherm 5.0 contains 17 113 entries from 771 proteins,

retrieved from 1497 scientific articles, which is ∼20% increase in data from the last version (4). Currently, the numbers of data for wild-type proteins, single, double and multiple mutants are 7014, 8202, 1277 and 620, respectively. Based on the solvent accessibility of mutants, 4426 mutations are buried, 2687 partially buried and 2751 exposed. In terms of secondary structures, 3993 mutations are in helix, 2622 in strand, 1227 in turn and 2467 in coil regions. Majority of data are obtained from CD (6825) and DSC (5294) experiments followed by fluorescence (3628). Further, 10 154 data are obtained by thermal denaturation, 3890 and 2796 data from GdnHCl and urea denaturation, respectively.

Currently, ProNIT 2.0 contains 4900 entries from 273 research papers. There are 158 different DNA-binding proteins with 3489 wild-type entries and 1411 mutant entries. Majority of data are obtained by gel shift (1316), fluorescence (1143) and filter binding (1053), followed by calorimetry (727), surface plasmon resonance (185) and footprinting (168). Although proteins from a variety of organisms are present in ProNIT, majority of interaction data are from *Escherichia coli* proteins (1625) followed by *Mus musculus* (637) and *Homo sapiens* (569).

NEW FEATURES

There is a growing interest in the relationship between structure and thermodynamics of proteins. Thus, we try to provide link from thermodynamic data in ProTherm to structural information. So far, ProTherm data are connected to sequence and structural information of proteins through 3DinSight. We have added a new cross-link between ProTherm and STING (9), a comprehensive analysis tool for proteins with many structural descriptors. For given protein mutations searched within STING, each entry in STING report is connected to available thermodynamic data in ProTherm. Conversely, all the ProTherm data with available protein structure have pointers to the corresponding STING entry with detailed structural information. This cross-link will greatly facilitate the analysis of structure–thermodynamic relationship of proteins. The ProTherm page also provides cross-reference tables necessary for creating cross-links with PDB (10), PIR (11) and Swiss-Prot (12) databases. We have also added several new features in the search interface to make the search more efficient and convenient.

In the current release of ProNIT 2.0, we have included ∼200 protein–RNA interaction data. To facilitate the retrieval of data based on DNA and RNA separately, we have added a new field called TYPE_NUC, where we provide the information about whether the nucleic acid sequence is DNA (single-stranded DNA, ssDNA, or double-stranded DNA, dsDNA) or RNA. Furthermore, a search option is added to retrieve data based on ssDNA, dsDNA or RNA. The protein nomenclature in the literature is not necessarily uniform. Hence, we have added a new field, SYNONYMS, in order to address this problem. Other additions of field are the SwissProt ID of the protein and 'RELATED_ENTRIES', which provides the list of entries that contain data from the same paper (the original paper usually contains multiple data and they are entered in different entries). We also provide a link to all homologous PDB codes with sequence identity of >95%. Also, display

options and sorting options are significantly improved. We have supplied the lists of ProNIT entries, protein names, protein sources, PDB code, NDB codes (13), authors and references in the advanced search page, along with a new query help page to help users for easy retrieval of the data. Several entries are deleted because of duplication, co-operative binding and so on, and the database entries are now renumbered. A mapping table, which relates the old and new entry numbers, is provided to help old users of ProNIT.

CITATION AND AVAILABILITY

The URLs for ProTherm and ProNIT are http://gibk26.bse.kyutech.ac.jp/jouhou/Protherm/protherm.html and http://gibk26.bse.kyutech.ac.jp/jouhou/pronit/pronit.html, respectively. The users of ProTherm and ProNIT are requested to cite the references (4) and (5), respectively, in their publication including the above URLs. Users who use both the databases for their work may cite this article in their publications. Suggestions and other materials for inclusion in the databases are welcome and should be sent to either protherm@rtcmain.bse.kyutech.ac.jp or pronit@rtcmain.bse.kyutech.ac.jp.

ACKNOWLEDGEMENTS

Conflict of interest statement. None declared.

REFERENCES

1. Gromiha,M.M., An,J., Kono,H., Oobatake,M., Uedaira,H. and Sarai,A. (1999) ProTherm: Thermodynamic database for proteins and mutants. *Nucleic Acids Res.*, **27**, 286–288.
2. Gromiha,M.M., Uedaira,H., An,J., Selvaraj,S., Prabakaran,P. and Sarai,A. (2002) ProTherm: thermodynamic database for proteins and mutants: developments in version 3.0. *Nucleic Acids Res.*, **30**, 301–302.
3. Sarai,A., Gromiha,M.M., An,J., Prabakaran,P., Selvaraj,S., Kono,H., Oobatake,M. and Uedaira,H. (2002) Thermodynamic databases for proteins and protein–nucleic acid interactions. *Biopolymers*, **61**, 121–126.
4. Abdulla Bava,K., Gromiha,M.M., Uedaira,H., Kitajima,K. and Sarai,A. (2004) ProTherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res.*, **32**, D120–D121.
5. Prabakaran,P., An,J., Gromiha,M., Selvaraj,S., Uedaira,H., Kono,H. and Sarai,A. (2001) Thermodynamic database for protein–nucleic acid interactions (ProNIT). *Bioinformatics*, **17**, 1027–1034.
6. Pfeil,W. (1998) *Protein Stability and Folding: A Collection of Thermodynamic Data*. Springer, NY.
7. Pfeil,W. (2001) *Protein Stability and Folding, Supplement1: A Collection of Thermodynamic Data*. Springer, NY.
8. An,J., Nakama,T., Kubota,Y. and Sarai,A. (1998) 3DinSight: an integrated relational database and search tool for structure, function and property of biomolecules. *Bioinformatics*, **14**, 188–195.
9. Neshich,G., Borro,L.C., Higa,R.H., Kuser,P.R., Yamagishi,M.E., Franco,E.H., Krauchenco,J.N., Fileto,R., Ribeiro,A.A., Bezerra,G.B. *et al.* (2005) The Diamond STING server. *Nucleic Acids Res.*, **33**, W29–W35.
10. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data bank. *Nucleic Acids Res.*, **28**, 235–242.
11. Cathy,H.W., Yeh,L.L., Huang,H., Arminski,L., Jorge,C.A., Chen,Y., Hu,Z.Z., Ledley,R.S., Kourtesis,P., Suzek,B.E. *et al.* (2003) The Protein Information Resource. *Nucleic Acids Res.*, **31**, 345–347.
12. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
13. Berman,H.M., Zardecki,C. and Westbrook,J. (1998) The Nucleic Acid Database: a resource for nucleic acid science. *Acta Crystallogr. D Biol. Crystallogr.*, **D54**, 1095–1104.