D192–D194 *Nucleic Acids Research, 2006, Vol. 34, Database issue*
doi:10.1093/nar/gkj005

CyBase: a database of cyclic protein sequence and structure

Jason P. Mulvenna, Conan Wang and David J. Craik*

Institute for Molecular Bioscience, University of Queensland, Brisbane, Queensland 4072, Australia

Received June 22, 2005; Revised and Accepted August 24, 2005

ABSTRACT

CyBase is a curated database and information source for backbone-cyclized proteins. The database incorporates naturally occurring cyclic proteins as well as synthetic derivatives, grafted analogues and acyclic permutants. The database provides a centralized repository of information on all aspects of cyclic protein biology and addresses issues pertaining to the management and searching of topologically circular sequences. The database is freely available at http://research.imb.uq.edu.au/cybase.

INTRODUCTION AND MOTIVATION

In recent years a number of proteins have been discovered that contain a macrocyclic backbone consisting of a continuous cycle of peptide bonds. Such macrocyclic proteins were unknown a decade ago but have now been discovered in bacteria, plants and animals (1). Unlike bacterial polyketides and small cyclic peptides such as cyclosporin that are constructed by peptide synthetases (2), these new cyclic proteins are ribosomally produced gene products, with backbone cyclization occurring as a post-translational modification. This new class of protein has excited interest because circular proteins have a range of advantages over conventional proteins (3,4). At least one class of cyclic proteins, the cyclotides, has been shown to be resistant to proteolysis and to a wide variety of adverse thermal and chemical conditions (5,6). Furthermore, since the termini of conventional proteins are often flexible, and as the degree of flexibility can be reduced by cyclization, entropic factors can lead to improved receptor binding affinities of circular proteins over corresponding acyclic proteins (7,8).

Five classes of naturally occurring proteins contain cyclic examples. These include the cyclic sex-pilin (9) and three bacteriocins (10–12) from bacterial sources, trypsin inhibitors from sunflower seeds (SFTI-1) (7,8) and the squash plant *Momordica cochinchinensis* (MCoTI-II) (13), the θ-defensins from macaque monkeys (14) and the cyclotides from plants of the Violaceae and Rubiaceae (15–17). The cyclotides are by far the largest family of circular proteins and ~60 cyclotide sequences have been reported thus far. Screening programs suggest that the number of sequences may soon number in the thousands (18,19). In addition to this natural diversity, a large number of synthetic mutants, grafted analogues and acyclic permutants of cyclic proteins have been reported, and several proteins of biological interest have been artificially cyclized (20–22). This growth in information necessitates a collation of sequence/structure/function data and the development of a uniform nomenclature to prevent duplication of research and multiple naming schemes. Furthermore, sequence searching of cyclic proteins adds an extra layer of complexity, with most sequence searching tools assuming a linear sequence of amino acids. Because backbone cyclization is a 'seamless' post-translational modification, the location of the N- and C-termini cannot be determined from the mature sequence alone. Consequently, when searching cyclic sequences, the point at which the sequence begins and ends in the database is often arbitrary and may confound traditional searching techniques. The special considerations needed for dealing with cyclic sequences and the rapidly expanding data on their structure and function has led us to develop a curated database and web-based information source for cyclic proteins called CyBase.

APPLICATION AND DISCUSSION

CyBase incorporates a MySQL database that contains a repository of information on the amino acid and nucleic acid sequence, structure and activity of cyclic proteins. The layout of the database is shown in Figure 1. At the core of the database is the protein table, which contains information on each cyclic protein characterized. Related to each protein table entry are tables containing information on nucleic acid sequences, structure, activity and literature references. A web-based interface provides access to the information and allows for text-based searching on all data fields and filtering of results by class, source, activity or other attributes. To account for the cyclic nature of the sequences any sequence search uses a concatenation of two copies of the linear representation of the sequence to simulate a cyclic protein. Each protein has

*To whom correspondence should be addressed. Tel: +61 7 3346 2019; Fax: +61 7 3346 2029; Email: d.craik@imb.uq.edu.au

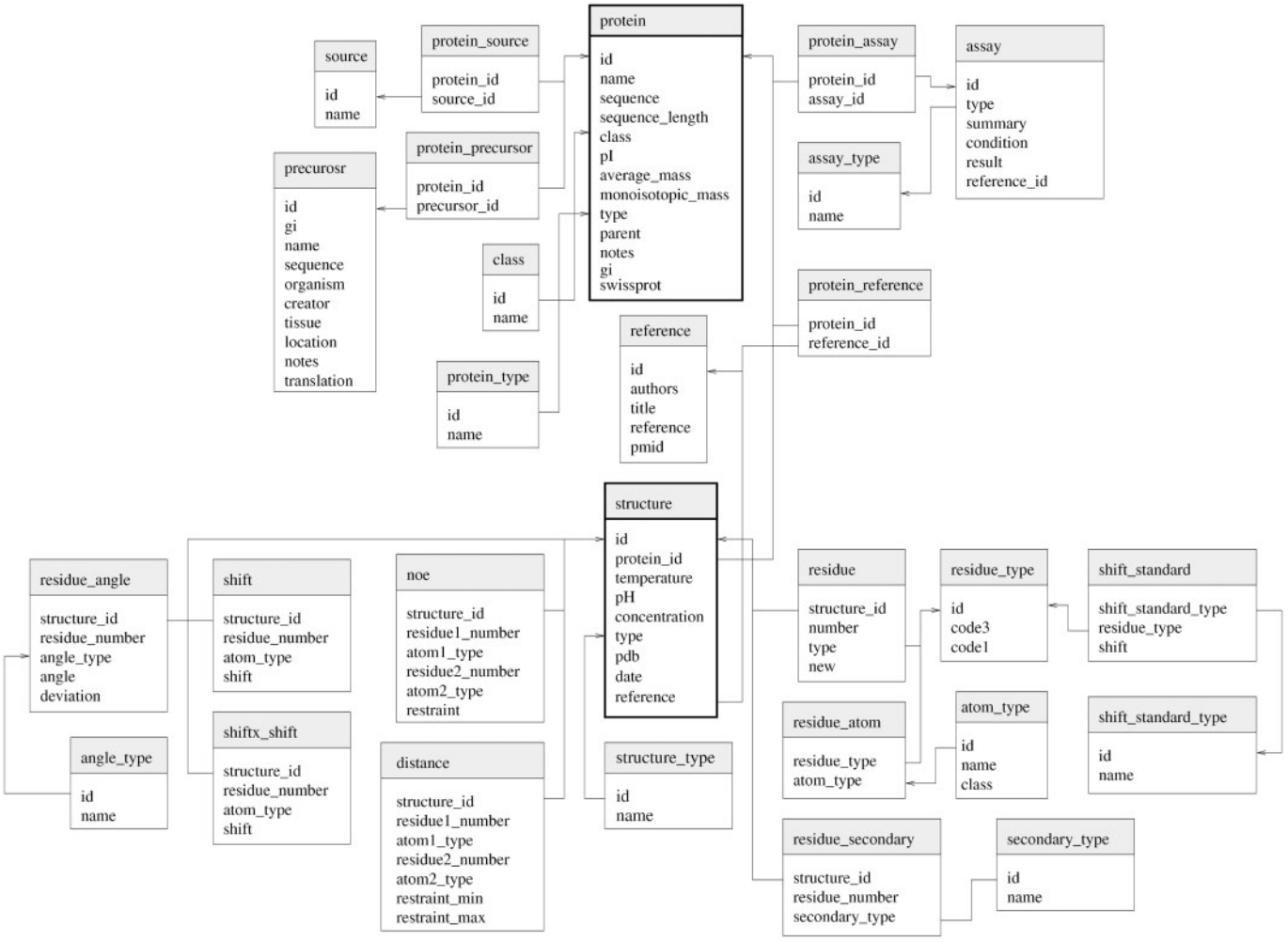

Figure 1. Schematic of the relational database underlying CyBase.

a dynamically produced sequence card, which provides cross-links to activity, nucleic acid sequence and structural information contained in the database. As the database is intended to supplement existing biological databases, links to UniProt Knowledgebase, Genbank and PDB are available for each entry and linkage with the KNOTTIN website (23) is planned. A number of other tools are provided, including coloured alignments and calculation of amino acid frequencies of selected sequences.

In general, naturally occurring cyclic proteins are small, with the largest possessing 78 amino acids. This small size makes the combination of mass spectrometry, to obtain sequence information, and NMR, to determine 3D structures, ideal for characterizing these proteins. Accordingly, to facilitate the rapid characterization of newly discovered proteins the database can be queried on molar mass, and for cyclotides, the capability exists for searching on the mass of fragments corresponding to particular inter-cysteine loops, facilitating sequence determination when utilizing reduction/alkylation of cysteine residues and tandem mass spectrometry (24,25).

Analysis of NMR-derived data such as chemical shifts and patterns of NOE connectivity can provide an early indication of the structure of a protein. To facilitate rapid structural characterization of newly discovered cyclic proteins chemical shift and restraint data from NMR-derived structures are included in the database, along with dihedral angle information. From these data, distances and regions containing defined secondary structure are calculated and stored in the database. These data can be presented visually for the analysis of short- and long-range NOE patterns, the backbone dihedral angles and chemical shift patterns. Although NMR is the most common technique for analysing these proteins, X-ray structures are also incorporated into the database and sets of inter-atom distances calculated for comparative purposes. As with the protein and nucleic acid entries, each structure possesses an information card, which contains cross-links to protein and nucleic acid entries.

Updating of the database is facilitated by a range of PERL and PHP scripts. These provide for the automated searching of sequence databases, using BLAST, to provide examples of novel cyclic proteins, and ensure quality control by preventing duplication of sequence data and renaming of already characterized sequences, a particularly important consideration for the cyclotide family, which contains potentially many sequences that may occur in a number of different species. These scripts also provide for the standardizing of residue

numbering in new cyclotide structures. Despite these automations the addition of new entries is performed manually to ensure maximum quality of database entries.

We plan to extend the database by utilizing the growing number of cyclotide structures to provide predictions of cyclotide secondary structures based on primary sequence and to develop methods to search the structures in the database based on the similarity between selected inter-atomic distances and NOE connectivities. We also plan to improve the information content of the database by including hydrogen bond and other structural information as well as homology models for cyclotide sequences that have not yet been structurally characterized. CyBase is available at http://research.imb.uq.edu.au/cybase/ and given the growing interest in backbone cyclization, it is hoped that CyBase will prove to be a useful resource in the field of structural biology. Suggestions should be directed to D.C.

ACKNOWLEDGEMENTS

Development of CyBase was supported by a grant from the Australian Research Council. Funding to pay the Open Access publication charges for this article was provided by D. Craik.

Conflict of interest statement. None declared.

REFERENCES

1. Trabi,M. and Craik,D. (2002) Circular proteins: no end in sight. *Trends Biochem. Sci*., **27**, 132–138.
2. Kohli,R. and Walsh,C. (2003) Enzymology of acyl chain macrocyclization in natural product biosynthesis. *Chem. Commun. (Camb.)*, **3**, 297–307.
3. Zhou,H. (2004) Loops, linkages, rings, catenanes, cages, and crowders: entropy-based strategies for stabilizing proteins. *Acc. Chem. Res*., **37**, 123–130.
4. Felizmenio-Quimio,M., Daly,N. and Craik,D. (2001) Circular proteins in plants—solution structure of a novel macrocyclic trypsin inhibitor from *Momordica cochinchinensis*. *J. Biol. Chem*., **276**, 22875–22882.
5. Gran,L., Sandberg,F. and Sletten,K. (2000) Oldenlandia affinis (R&S) DC. A plant containing uteroactive peptides used in African traditional medicine. *J. Ethnopharmacol*., **70**, 197–203.
6. Colgrave,M. and Craik,D. (2004) Thermal, chemical, and enzymatic stability of the cyclotide kalata B1: the importance of the cyclic cystine knot. *Biochemistry*, **43**, 5965–5975.
7. Luckett,S., Garcia,R., Barker,J., Konarev,A., Shewry,P., Clarke,A. and Brady,R. (1999) High resolution structure of a potent, cyclic proteinase inhibitor from sunflower seeds. *J. Mol. Biol*., **290**, 525–533.
8. Korsinczky,M., Schirra,H., Rosengren,K., West,J., Condie,B., Otvos,L., Anderson,M. and Craik,D. (2001) Solution structures by 1H NMR of the novel cyclic trypsin inhibitor SFTI-1 from sunflower seeds and an acyclic permutant. *J. Mol. Biol*., **311**, 579–591.
9. Eisenbrandt,R., Kalkum,M., Lai,E., Lurz,R., Kado,C. and Lanka,E. (1999) Conjugative pili of IncP plasmids, and the Ti plasmid T pilus are composed of cyclic subunits. *J. Biol. Chem*., **274**, 22548–22555.
10. Kawai,Y., Saito,T., Toba,T., Samant,S. and Itoh,T. (1994) Isolation and characterization of a highly hydrophobic new bacteriocin (gassericin A) from *Lactobacillus gasseri* LA39. *Biosci. Biotechnol. Biochem*., **58**, 1218–1221.
11. Kemperman,R., Kuipers,A., Karsens,H., Nauta,A., Kuipers,O. and Kok,J. (2003) Identification and characterization of two novel clostridial bacteriocins, circularin A and closticin 574. *Appl. Environ. Microbiol*., **69**, 1589–1597.
12. Maqueda,M., Galvez,A., Bueno,M., Sanchez-Barrena,M., Gonzalez,C., Albert,A., Rico,M. and Valdivia,E. (2004) Peptide AS-48: prototype of a new class of cyclic bacteriocins. *Curr. Protein. Pept. Sci*., **5**, 399–416.
13. Hernandez,J., Gagnon,J., Chiche,L., Nguyen,T., Andrieu,J., Heitz,A., Hong,T.T., Pham,T. and Nguyen,D.L. (2000) Squash trypsin inhibitors from *Momordica cochinchinensis* exhibit an atypical macrocyclic structure. *Biochemistry*, **39**, 5722–5730.
14. Tang,Y., Yuan,J., Osapay,G., Osapay,K., Tran,D., Miller,C., Ouellette,A. and Selsted,M. (1999) A cyclic antimicrobial peptide produced in primate leukocytes by the ligation of two truncated θ-defensins. *Science*, **286**, 498–502.
15. Craik,D., Daly,N., Bond,T. and Waine,C. (1999) Plant cyclotides: a unique family of cyclic and knotted proteins that defines the cyclic cystine knot structural motif. *J. Mol. Biol*., **294**, 1327–1336.
16. Jennings,C., West,J., Waine,C., Craik,D. and Anderson,M. (2001) Biosynthesis and insecticidal properties of plant cyclotides: the cyclic knotted proteins from *Oldenlandia affinis*. *Proc. Natl Acad. Sci. USA*, **98**, 10614–10619.
17. Goransson,U., Svangard,E., Claeson,P. and Bohlin,L. (2004) Novel strategies for isolation and characterization of cyclotides: the discovery of bioactive macrocyclic plant polypeptides in the Violaceae. *Curr. Protein Pept. Sci*., **5**, 317–329.
18. Craik,D., Daly,N., Mulvenna,J., Plan,M. and Trabi,M. (2004) Discovery, structure and biological activities of the cyclotides. *Curr. Protein Pept. Sci*., **5**, 297–315.
19. Gustafson,K., McKee,T. and Bokesch,H. (2004) Anti-HIV cyclotides. *Curr. Protein Pept. Sci*., **5**, 331–340.
20. Daly,N. and Craik,D. (2000) Acyclic permutants of naturally occurring cyclic proteins. Characterization of cystine knot and β-sheet formation in the macrocyclic polypeptide kalata B1. *J. Biol. Chem*., **275**, 19068–19075.
21. Deechongkit,S. and Kelly,J. (2002) The effect of backbone cyclization on the thermodynamics of beta-sheet unfolding: stability optimization of the PIN WW domain. *J. Am. Chem. Soc*., **124**, 4980–4986.
22. Williams,N., Prosselkov,P., Liepinsh,E., Line,I., Sharipo,A., Littler,D., Curmi,P., Otting,G. and Dixon,N. (2002) *In vivo* protein cyclization promoted by a circularly permuted *Synechocystis sp*. PCC6803 DnaB mini-intein. *J. Biol. Chem*., **277**, 7790–7798.
23. Gelly,J., Gracy,J., Kaas,Q., Le-Nguyen,D., Heitz,A. and Chiche,L. (2004) The KNOTTIN website and database: a new information system dedicated to the knottin scaffold. *Nucleic Acids Res*., **32**, D156–D159.
24. Goransson,U. and Craik,D. (2003) Disulfide mapping of the cyclotide kalata B1. Chemical proof of the cystic cystine knot motif. *J. Biol. Chem*., **278**, 48188–48196.
25. Mulvenna,J., Sando,L. and Craik,D. (2005) Processing of a 22 kDa precursor protein to produce the circular protein tricyclon A. *Structure*, **13**, 691–701.