D314–D319 *Nucleic Acids Research, 2008, Vol. 36, Database issue* *Published online 19 November 2007*
doi:10.1093/nar/gkm939

KNOTTIN: the knottin or inhibitor cystine knot scaffold in 2007

Jérôme Gracy[1,2], Dung Le-Nguyen[3], Jean-Christophe Gelly[4], Quentin Kaas[5], Annie Heitz[1,2] and Laurent Chiche[1,2,*]

[1]Université de Montpellier, CNRS, UMR5048, Centre de Biochimie Structurale, 34090 Montpellier, [2]INSERM, U554, Montpellier, [3]CNRS, FRE3009; BIORAD; Complex system modelling and engineering for diagnostic, Montpelllier, [4]CNRS, UPR 9080, Université Paris 7, Laboratoire de Biochimie Théorique, Institut de Biologie Physico-Chimique, Paris, France and [5]Institute for Molecular Bioscience, The University of Queensland, Brisbane, Australia

Received September 14, 2007; Revised October 8, 2007; Accepted October 11, 2007

ABSTRACT

The KNOTTIN database provides standardized information on the small disulfide-rich proteins with a knotted topology called knottins or inhibitor cystine knots. Static pages present the essential historical or recent results about knottin discoveries, sequences, structures, syntheses, folding, functions, applications and bibliography. New tools, KNOTER3D and KNOTER1D, are provided to determine or predict if a user query (3D structure or sequence) is a knottin. These tools are now used to automate the database update. All knottin structures and sequences in the database are now standardized according to the knottin nomenclature based on loop lengths between knotted cysteines, and to the knottin numbering scheme. Therefore, the whole KNOTTIN database (sequences and structures) can now be searched using loop lengths, in addition to keyword and sequence (BLAST, HMMER) searches. Renumbered and structurally fitted knottin PDB files are available for download as well as renumbered sequences, sequence alignments and logos. The knottin numbering scheme is used for automatic drawing of standardized two-dimensional Colliers de Perles of any knottin structure or sequence in the database or provided by the user. The KNOTTIN database is available at http://knottin.cbs.cnrs.fr.

KNOTTINS: SMALL DISULFIDE-RICH PROTEINS WITH A KNOTTED ARRANGEMENT

The knottins are fascinating miniproteins present in many species and featuring various biological actions such as toxic, inhibitory, anti-microbial, insecticidal, cytotoxic, anti-HIV or hormone-like activity (1). They share a unique knotted topology of three disulfide bridges, with one disulfide penetrating through a macrocycle formed by the two other disulfides and inter-connecting peptide backbones. This scaffold was first discovered in 1982 in PCI, a carboxypeptidase inhibitor from potato (2). It was since observed in an amazing number of unrelated protein families including, e.g. toxins from plants, bugs, molluscs or arachnids, or anti-microbials from plants, insects or arthropods. The KNOTTIN database provides standardized data on sequences, structures and other information on known knottins (3). Proteins sharing this scaffold were referred to as knottins (4) or inhibitor cystine knots (ICK) (5), or even simply as cystine knots. The most populated knottin families are conotoxins (389 sequences), spider toxins (257 sequences) and cyclotides (105 sequences). The cyclotides are head-to-tail cyclized knottins present in plants from the Violaceae and Rubiaceae [see www.cyclotide.com (6)]. These miniproteins are considered as natural combinatorial peptide libraries structurally constrained by the knottin scaffold (7,8) but in which hypermutation of essentially all residues are permitted with the exception of the strictly conserved cysteines of the knot. The main knottin features are therefore a remarkable stability due to the cystine knot, a small size making them readily accessible to chemical synthesis, and an excellent tolerance to sequence variations. Knottins thus appear as appealing leads or frameworks for peptide drug design (1,9–18). One knottin has come to market for the treatment of chronic pain (Prialt from Elan corp http://www.elan.com) and others are on the way. Companies have emerged that plan to use knottins as leads or scaffolds in drug design (e.g. see the 'Microbody' technology, i.e. generation of drug candidates based on knottins, at www.nascacell.de).

The new developments in the field have prompted us to improve the KNOTTIN database content. Moreover, following collaboration with the Swiss Institute of Bioinformatics, UniProtKB/Swiss-Prot entries (19)

*To whom correspondence should be addressed. Tel: +(33) 4 67 41 77 03; Fax: +(33) 4 67 41 79 13; Email: laurent.chiche@cbs.cnrs.fr

Table 1. Statistics on the current KNOTTIN database content

Family	Cys IV	1D[a]	NMR[b]	X-ray[c]	Organisms	
					1D[a]	3D[a]
All		1066	126	14	299	61
Agouti-related	61	95	4		75	2
Alpha-amylase inhibitor	61	1	2	1	1	1
Bug	61	3	2		3	2
Carboxypeptidase inhibitor	77	13	1	1	4	1
Conotoxin1	61	389	26		52	10
Conotoxin2	78	4	1		4	1
Conotoxin3	77	1			1	
Cyclotide	78	105	20		20	8
Fungi1	77	4			4	
Fungi2	61	2			2	
Gurmarin like	61	1	2		1	1
Horseshoe crab	61	4	3		1	1
Insect antimicrobial	61	13	1		6	1
Phenoloxidase inhibitor	61	5			4	
Plant antimicrobial	61	6	1		3	1
Plant defensin	63	2	1		1	1
Plant toxin	78	30	2		17	2
Scorpion1	63	14	2		7	2
Scorpion2	61	5	2		4	2
Scorpion3	63	25	2		6	1
Spider	61	257	44		47	15
Sponge	61	1			1	
Terebra	61	1			1	
Serine protease inhibitor1	78	35	8	12	16	8
Serine protease inhibitor2	61	4			2	
Trematoda	61	15			1	
Virus1	61	22	2		6	1
Virus2	61	9			9	

[a]1D refers to the sequence database, 3D to the structure database.
[b]Number of NMR structures.
[c]Number of X-ray structures.

are now annotated with knottin structural information. Relevant entries can be retrieved with the newly introduced keyword 'Knottin'.

THE KNOTTIN DATABASE IN 2007

The first database release (3) contained 11 protein families, 85 three-dimensional (3D) structures and 385 sequences (1D). The content has more than doubled in the current release built from the Uniprot release 11.3 [10 July 2007; UniProtKB/Swiss-Prot 53.3 and UniProtKB/TrEMBL 36.3 (19)] using new automated tools (see subsequently). The KNOTTIN database now contains 28 families, 145 3D structures and 1066 protein sequences. Table 1 displays an overview of the current release content.

Several families have no 3D structures included in the database. For two of them, Fungi1 and Sponge, 3D structures have been reported that unambiguously classify these proteins as knottins (20,21) although these structures were not deposited into the Protein Data Bank (22). Nevertheless, seven families were included despite the fact that no 3D structure has yet been reported. These families, Conotoxin3, Fungi2, Phenoloxidase inhibitor, Serine protease inhibitor2, Terebra, Trematoda and Virus2, were selected for inclusion because all of them displayed good similarity with known knottins according to the KNOTER1D prediction tool (see subsequently). Some of them were moreover previously predicted as knottins [Phenoloxidase inhibitor (23)], have known disulfide bridges [Serine protease inhibitor2 (24)], were classified by others in protein families already included in the KNOTTIN database (Conotoxin3 is classified in the conotoxin-P family; Virus2 is classified in the conotoxin-O superfamily) or were annotated as knottins by the UniProt depositors (Trematoda). Finally, the Fungi2 and the Terebra families were included purely based on the KNOTER1D predictions and should probably be viewed with caution.

Other improvements of the KNOTTIN website are, e.g. the availability of alignments in various text formats for downloads and of sequence logos (25). The logo for the entire database content is shown in Figure 1 and the percent of identity between knottins in Figure 2.

The textual content of the database (menus 'Functions', 'Folding', 'Synthesis', 'Modeling & drug design', 'Landmarks', 'References', 'Links') has been regularly updated. Its display has been improved and reorganized, and the cited references are now included in the database thus allowing keyword searches. It is not expected to provide exhaustive data on knottins, but rather to permanently provide a freely available review of essential data on knottins.

The KNOTTIN database is freely available at http://knottin.cbs.cnrs.fr or http://knottin.com, but we request that this article be cited when using the KNOTTIN database in research projects.

AUTOMATED DETECTION OF KNOTTINS: KNOTER3D AND KNOTER1D

Previous updates of the KNOTTIN database implied numerous manual steps, e.g. manual inspection of BLAST (26), PSI-BLAST (27) or HMMER (28) search results for each knottin family. To (i) facilitate and speed up the database update, (ii) reduce unavoidable errors due to manual steps and (iii) provide prediction tools for new sequences or structures, we have implemented two new algorithms named KNOTER3D and KNOTER1D that largely automate knottin detection.

Given a protein structure Str1, KNOTER3D first searches for the presence of three disulfide bridges with the I–IV, II–V, III–VI connectivity. If present, the Str1 protein in renumbered such that cysteines I, II, III, V and VI have numbers 20, 40, 60 80 and 100, respectively. Then the structural core of Str1, i.e. the cystine-stabilized beta-sheet motif (29) (renumbered residues 40, 60–61, 79–81 and 99–100), is superimposed onto the corresponding motif of a reference knottin structure (CPTI-II, PDB ID: 2btcI) and a RMSD below 2.5 Å indicates that the structure Str1 is a knottin.

The automated detection of knottins from their amino acid sequence alone is more difficult. The KNOTER1D prediction tool is based on sequence similarity search, cysteine position analysis and, when possible, database

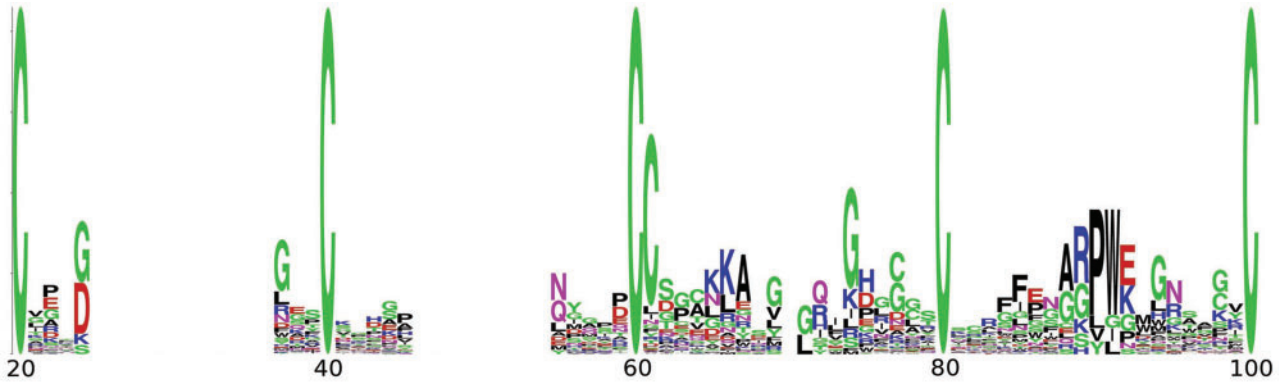

Figure 1. Sequence logo (25) for the alignment of all standardized sequences in the KNOTTIN database. The alignment is truncated between the first and the last cysteine of the knot (standard positions 20 and 100, respectively).

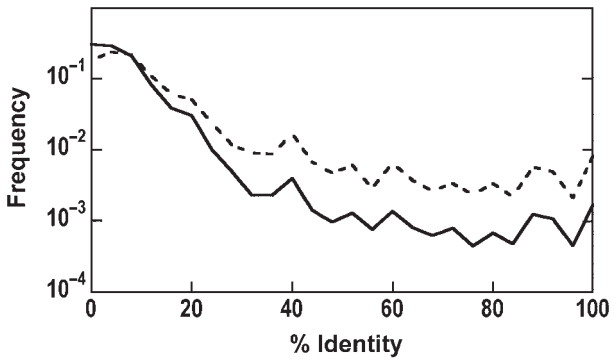

Figure 2. Frequencies of percent identity observed when comparing each knottin with all other knottins (plain line) or with other knottins in the same family (dashed line).

annotation mining if a close homolog of the considered protein is available in the UniProt database.

Given a protein sequence Seq1 whose the knottin status is unknown, the prediction algorithm consists in searching sequences similar to Seq1 from the KNOTTIN database using BLAST, then comparing Seq1 with each found similar knottin sequence Knot_Seq2 and its associated knottin family Knot_Fam2 until the following similarity score is above a predefined cutoff. The similarity between Seq1 and Knot_Seq2 is measured using a composite score integrating the following sequence analysis data:

- S is a sequence similarity score. It is the *P*-value logarithm of the best local pair-wise alignment between Seq1 and Knot_Seq2 detected by BLAST.
- C is a score related to knotted cysteines. It is the number of query cysteines, which correspond to knotted cysteines when aligning Seq1 onto the multiple sequence alignment of family Knot_Fam2 using CLUSTALW.
- L is a loop compatibility score. It is computed using a first-order hidden Markov model based on inter-cysteine loop-length frequencies observed in family Knot_Fam2.
- Furthermore, if Uni_Seq1, the Uniprot database entry whose amino acid sequence is the most similar to Seq1 according to BLAST, shares more than 80% sequence identity with Seq1, then the previously defined scores S, C and L are complemented with the following database-mining pieces of information. The information can also be provided directly to KNOTER1D.
- T is a taxonomic score. A specific and a general taxonomic group are defined for each knottin family, e.g. 'Cucurbitaceae' and 'Viridiplantae', respectively, for the 'Serine protease Inhibitor1' family. Then, T = 2 if the Uni_Seq1 species belongs to the specific group of the knottin family Knot_Fam2, T = 1 if the Uni_seq1 species belongs to the general group only and T = 0 if the Uni_seq1 species does not belong to the general taxonomic group of the knottin family Knot_Fam2.
- F is a functional score. A list of functions is defined for each knottin families. Then F = 1 if the Uni_Seq1 entry function annotation is compatible with any function of the knottin family Knot_Fam2 (e.g. 'Protease inhibitor' for the Serine Protease Inhibitor1 family).
- K is a score based on keywords. A list of keywords was established from data in the DE, KW, DR, FT and CC fields of knottins and other disulfide-rich proteins contained in a disulfide-rich subset of the UniProt/Swiss-Prot database. Associated weights were empirically set to positive or negative values according to the ratio between knottin or non-knottin proteins in the subset of UniProt/Swiss-Prot entries matching the considered keyword (e.g. the keyword 'violacin' which matches knottin entries only has a weight set to +6, while the keyword 'egf' which defines non-knotted cysteine-rich modules has a weight set to –8).

Otherwise, if Uni_Seq1 is missing in Uniprot, i.e. there is no close homolog of the query sequence in UniProt, then the scores T, F and K are set to 0.

A composite knottin similarity score is then built as the weighted sum

$$TS = wS^{*}S + wC^{*}C + wL^{*}L + wT^{*}T + wF^{*}F + wK^{*}K.$$

A subset of UniProt/SwissProt + TrEMBL containing only disulfide-rich proteins [SS(UniProt/SP-Tr), 137 875 proteins] has been generated. Then weights wS, wC, wL, wT, wF, wK and the score cutoff for knottin prediction were optimized against SS(UniProt/SP-Tr), in order to maximize the discrimination between knottins and non-knottins according to the current release of the KNOTTIN database. Based on the TS composite score, KNOTER1D predicts the query sequence Seq1 as a knottin, a putative knottin, or a non-knottin.

For knottin or putative knottin, KNOTER1D provides:

- A multiple alignment of the query sequence onto the closest knottin family.
- The sequence number of the six cysteines predicted to be involved in the knot.
- A renumbered alignment according to the standard knottin numbering.
- The database update is now essentially based on KNOTER3D and KNOTER1D. All new structures in the Protein Data Bank (22) since the last update are submitted to the KNOTER3D tool and new detected knottins are integrated in the database. On the other hand, each sequence in SS[UniProt/SP-Tr] is sent to KNOTER1D and synthetic results are compiled in a list containing one line by sequence ranked by decreasing values of the composite score TS. Each protein line indicates the TS score, the query accession number, the query knottin family if it is already in the database, the most similar (hit) knottin ID and family, the BLAST score for query/hit alignment, the T, F, C and L scores, the total number of cysteines in the query, the length of the query and the keywords/weights used (K score).
- This file is first manually annotated to select or reject new sequences to be included in the KNOTTIN database, and then submitted to perl scripts that generate the updated 'sequence' section of the database. This mostly automated protocol significantly improves the sensitivity and reliability of knottin detection. It also helped in discovering new putative knottin families (Table 1).

KNOTTIN STANDARDIZATIONS

The knottin standardization relies only on the detection of the six cysteines of the knot (1,3). Briefly, the cysteines I, II, III, V and VI involved in the knot are renumbered 20, 40, 60, 80 and 100. The cysteine IV, which has different spatial locations in various families, receives different numbers in the range 61, e.g. in 'Conotoxin1' and in 'Spider toxins', to 78, e.g. in 'Serine protease inhibitor1' (Table 1). Although this is straightforward for knottins containing only six cysteines, it can become rather cumbersome for knottins containing more than six cysteines. The only way to rigorously and systematically determine the knotted cysteines is from 3D structure analyses. This is done by the KNOTER3D tool available in the Tool menu. However, by similarity, the standardization can then be transferred to any knottin sequence.

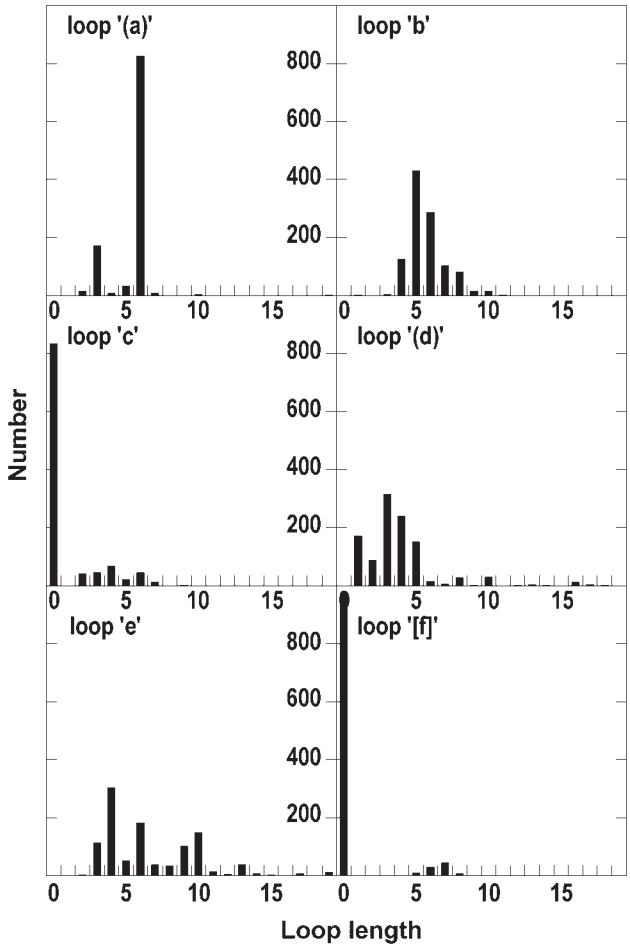

Figure 3. Statistics on loop lengths in the KNOTTIN database. Loop labels follow the knottin nomenclature (3).

This is now done by the KNOTER1D program (see above) that provides the knottin standardization for any sequence predicted to be a putative knottin.

Extension of the knottin standardization to all sequences in the current release has two main advantages:

(a) The sequence database can now be searched or sorted using criteria based on loop lengths. This allows, e.g. to easily establish exhaustive statistics on loop lengths including different unrelated families. Such statistics are displayed in Figure 3 and now appear on the database homepage. In previous releases, only the 3D structures were standardized.
(b) Standardized alignments are easily produced in which spatially conserved cysteines of the knot can be recognized (numbers 20, 40, 60, 80, 100) and are correctly aligned. Without such standardization, automatic alignments of non-homologous knottins are likely to be wrong. An example of a standardized alignment is shown in Figure 4 for knottins selected in the 'Insect anti-microbial' family.

As shown in Figure 3, all loops do not display similar length variability. Loops 'b', '(d)' and 'e' display rather

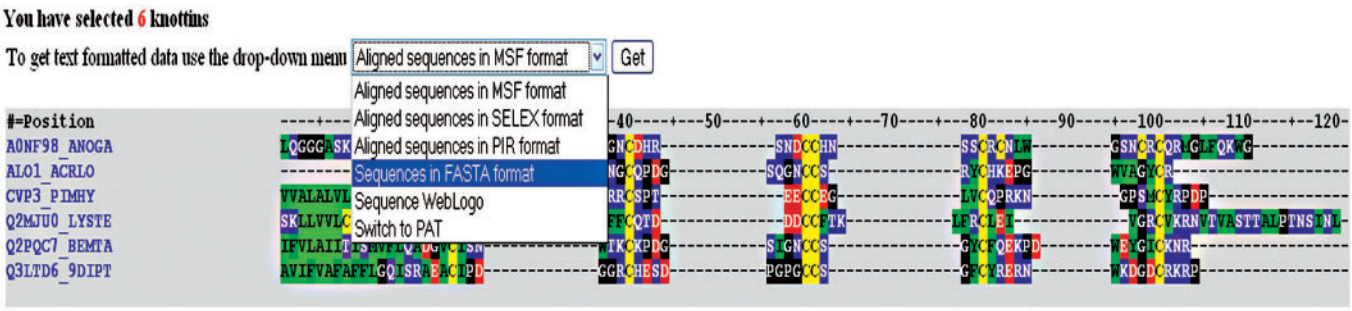

Figure 4. Standard alignment of selected knottins in the 'Insect anti-microbial' family. Cysteines numbered 20, 40, 60, 80 and 100 correspond to cysteines I, II, III, V and VI, respectively. The drop-down menu provides access to various alignment formats useful for data downloads. It also allows the creation of a logo from the alignment and the transfer of the alignment to the PAT webserver (30).

normal distributions although loop 'b' tends to be shorter and loop 'e' longer (it is worth noting that the loop 'e' hairpin often displays an additional stabilizing disulfide bridge between β-strands, see positions 82 and 98 in Figure 4). Loop '[f]' is the C-to-N linker in cyclic knottins, mainly cyclotides, and is shown as zero length for acyclic knottins. In contrast, loop '(a)' and 'c' display strongly biased lengths. Loop 'c' of zero length, i.e. cysteine IV is adjacent to cysteine III and is thus renumbered as 61 (Table 1), is present in Conotoxin1 and spider toxins, the two largest knottin families. Loop '(a)' also displays a striking distribution biased toward 3 and 6 lengths. These length biases remain to be explained.

The renumbered alignment shown in Figure 4 shows the new drop-down menu, which is useful to display text formatted data, but also to draw sequence logos or to send the alignment to the PAT webserver (30) allowing more powerful sequence analyses.

The KNOTER1D and KNOTER3D tools are available through the 'Tool' menu of the KNOTTIN database and through the 'Sequence similarity search' and 'Tertiary structure analysis' menus of the PAT webserver (30). The PAT webserver allows submission of protein sequences or structures in many formats including PDB or SwissProt IDs, and combinations with many other tools, while the KNOTTIN database provides 'Collier de Perles' standardized two-dimensional representations.

The standardization of all knottin sequences provided in the current release is the first main step toward rigorous 3D modeling of all knottins. Reconstruction of all atom models will consist then in loop modeling of inter-cysteine segments. A strategy for optimal reconstruction is currently underway.

ACKNOWLEDGEMENTS

This work was supported by the Centre National de la Recherche Scientifique (CNRS), the Institut National de la Santé et de la Recherche Médicale (INSERM). We thank Florence Jungo for fruitful discussions, Jean-Luc Pons for help in web developments and Marie-Paule Lefranc, for authorizing the use of the expression 'Collier de Perles' which originally refers to standardized 2D representations in IMGT, the international ImMunoGenetics information system® (http://imgt.cines.fr). Funding to pay the Open Access publication charges for this article was provided by the CNRS and the INSERM.

Conflict of interest statement. None declared.

REFERENCES

1. Chiche,L., Heitz,A., Gelly,J.C., Gracy,J., Chau,P.T., Ha,P.T., Hernandez,J.F. and Le-Nguyen,D. (2004) Squash inhibitors: from structural motifs to macrocyclic knottins. *Curr. Protein Pept. Sci.*, **5**, 341–349.
2. Rees,D.C. and Lipscomb,W.N. (1982) Refined crystal structure of the potato inhibitor complex of carboxypeptidase A at 2.5 A resolution. *J. Mol. Biol.*, **160**, 475–498.
3. Gelly,J.C., 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.
4. Le Nguyen,D., Heitz,A., Chiche,L., Castro,B., Boigegrain,R.A., Favel,A. and Coletti-Previero,M.A. (1990) Molecular recognition between serine proteases and new bioactive microproteins with a knotted structure. *Biochimie*, **72**, 431–435.
5. Pallaghy,P.K., Nielsen,K.J., Craik,D.J. and Norton,R.S. (1994) A common structural motif incorporating a cystine knot and a triple-stranded beta-sheet in toxic and inhibitory polypeptides. *Protein Sci.*, **3**, 1833–1839.
6. Wang,C.K.L., Kass,Q., Chiche,L. and Craik,D.J. (2007) CyBase: a database of cyclic protein sequences and structures, with applications in protein discovery and engineering. *Nucleic Acids Res*, doi:10.1093/nar/gkm953.
7. Sollod,B.L., Wilson,D., Zhaxybayeva,O., Gogarten,J.P., Drinkwater,R. and King,G.F. (2005) Were arachnids the first to use combinatorial peptide libraries? *Peptides*, **26**, 131–139.
8. Craik,D.J., Cemazar,M., Wang,C.K. and Daly,N.L. (2006) The cyclotide family of circular miniproteins: nature's combinatorial peptide template. *Biopolymers*, **84**, 250–266.
9. Krause,S., Schmoldt,H.U., Wentzel,A., Ballmaier,M., Friedrich,K. and Kolmar,H. (2007) Grafting of thrombopoietin-mimetic peptides into cystine knot miniproteins yields high-affinity thrombopoietin antagonists and agonists. *FEBS J.*, **274**, 86–95.
10. Clark,R.J., Daly,N.L. and Craik,D.J. (2006) Structural plasticity of the cyclic-cystine-knot framework: implications for biological activity and drug design. *Biochem. J.*, **394**, 85–93.
11. Craik,D.J., Cemazar,M. and Daly,N.L. (2006) The cyclotides and related macrocyclic peptides as scaffolds in drug design. *Curr. Opin. Drug Discov Dev.*, **9**, 251–260.
12. Hosse,R.J., Rothe,A. and Power,B.E. (2006) A new generation of protein display scaffolds for molecular recognition. *Protein Sci.*, **15**, 14–27.
13. Reiss,S., Sieber,M., Oberle,V., Wentzel,A., Spangenberg,P., Claus,R., Kolmar,H. and Losche,W. (2006) Inhibition of platelet

aggregation by grafting RGD and KGD sequences on the structural scaffold of small disulfide-rich proteins. *Platelets*, **17**, 153–157.
14. Werle,M., Schmitz,T., Huang,H.L., Wentzel,A., Kolmar,H. and Bernkop-Schnurch,A. (2006) The potential of cystine-knot microproteins as novel pharmacophoric scaffolds in oral peptide drug delivery. *J. Drug Target*, **14**, 137–146.
15. Binz,H.K., Amstutz,P. and Pluckthun,A. (2005) Engineering novel binding proteins from nonimmunoglobulin domains. *Nat. Biotechnol.*, **23**, 1257–1268.
16. Souriau,C., Chiche,L., Irving,R. and Hudson,P. (2005) New binding specificities derived from Min-23, a small cystine-stabilized peptidic scaffold. *Biochemistry*, **44**, 7143–7155.
17. Borman,S. (2004) Tying up loose ends: new examples and applications of circular and knotted peptides and proteins are turning up. *Chem. Eng. News*, **82**, 40–42.
18. Craik,D.J., Daly,N.L. and Waine,C. (2001) The cystine knot motif in toxins and implications for drug design. *Toxicon*, **39**, 43–60.
19. The UniProt Consortium (2007) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **35**, D193–D197.
20. Vervoort,J., van den Hooven,H.W., Berg,A., Vossen,P., Vogelsang,R., Joosten,M.H. and de Wit,P.J. (1997) The race-specific elicitor AVR9 of the tomato pathogen *Cladosporium fulvum*: a cystine knot protein. Sequence-specific 1H NMR assignments, secondary structure and global fold of the protein. *FEBS Lett.*, **404**, 153–158.
21. Takada,K., Hamada,T., Hirota,H., Nakao,Y., Matsunaga,S., van Soest,R.W. and Fusetani,N. (2006) Asteropine A, a sialidase-inhibiting conotoxin-like peptide from the marine sponge *Asteropus simplex*. *Chem. Biol.*, **13**, 569–574.
22. 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.
23. Daquinag,A.C., Sato,T., Koda,H., Takao,T., Fukuda,M., Shimonishi,Y. and Tsukamoto,T. (1999) A novel endogenous inhibitor of phenoloxidase from *Musca domestica* has a cystine motif commonly found in snail and spider toxins. *Biochemistry*, **38**, 2179–2188.
24. Kowalska,J., Pszczola,K., Wilimowska-Pelc,A., Lorenc-Kubis,I., Zuziak,E., Lugowski,M., Legowska,A., Kwiatkowska,A., Sleszynska,M. *et al.* (2007) Trypsin inhibitors from the garden four o'clock (*Mirabilis jalapa*) and spinach (*Spinacia oleracea*) seeds: isolation, characterization and chemical synthesis. *Phytochemistry*, **68**, 1487–1496.
25. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
26. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
27. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
28. Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK.
29. Heitz,A., Le-Nguyen,D. and Chiche,L. (1999) Min-21 and min-23, the smallest peptides that fold like a cystine-stabilized beta-sheet motif: design, solution structure, and thermal stability. *Biochemistry*, **38**, 10615–10625.
30. Gracy,J. and Chiche,L. (2005) PAT: a protein analysis toolkit for integrated biocomputing on the web. *Nucleic Acids Res.*, **33**, W65–W71.