Biochimica et Biophysica Acta 1844 (2014) 181–190

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbapap

Prediction and characterization of cyclic proteins from sequences in three domains of life[☆]

Pradyumna Kedarisetti [a], Marcin J. Mizianty [a], Quentin Kaas [b], David J. Craik [b], Lukasz Kurgan [a,*]

[a] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, T6G 2V4, Canada
[b] Institute for Molecular Bioscience, University of Queensland, Brisbane, QLD, 4072, Australia

ARTICLE INFO

Article history:
Received 2 February 2013
Received in revised form 12 April 2013
Accepted 2 May 2013
Available online 10 May 2013

Keywords:
Cyclic protein
Cyclotide
Prediction
CyPred
Cyclic defensin
Bacteriocin
Trypsin inhibitor

ABSTRACT

Cyclic proteins (CPs) have circular chains with a continuous cycle of peptide bonds. Their unique structural traits result in greater stability and resistance to degradation when compared to their acyclic counterparts. They are also promising targets for pharmaceutical/therapeutic applications. To date, only a few hundred CPs are known, although recent studies suggest that their numbers might be substantially higher. Here we developed a first-of-its-kind, accurate and high-throughput method called CyPred that predicts whether a given protein chain is cyclic. CyPred considers currently well-represented CP families: cyclotides, cyclic defensins, bacteriocins, and trypsin inhibitors. Empirical tests demonstrate that CyPred outperforms commonly used alignment methods. We used CyPred to estimate the incidence of CPs and found ~3500 putative CPs among 5.7+ million chains from 642 fully sequenced proteomes from archaea, bacteria, and eukaryotes. The median number of putative CPs per species ranges from three for archaea proteomes to two for eukaryotes/bacteria, with 7% of archaea, 11% of bacterial, and 16% of eukaryotic proteomes having 10+ CPs. The differences in the estimated fractions of CPs per proteome are as large as three orders of magnitude. Among eukaryotes, animals have higher ratios of CPs compared to fungi, while plants have the largest spread of the ratios. We also show that proteomes enriched in cyclic proteins evolve more slowly than proteomes with fewer cyclic chains. Our results suggest that further research is needed to fully uncover the scope and potential of cyclic proteins. A list of putative CPs and the CyPred method are available at http://biomine.ece.ualberta.ca/CyPred/. This article is part of a Special Issue entitled: Computational Proteomics, Systems Biology & Clinical Implications. Guest Editor: Yudong Cai.

1. Introduction

Cyclic proteins (CPs) have their termini linked together to create a cyclic backbone and thus effectively have no beginning and no end in their native conformation. Naturally produced circular proteins have been found in bacteria, plants, fungi, and animals [1,2]. Compared to their non-cyclic counterparts, they are relatively short (about a dozen to 100 amino acids), less prone to degradation, more structurally stable, and are harder to denature [1,3]. One of the largest CP families, cyclotides, comprise disulfide-rich chains of 28 to 37 amino acids with a characteristic cyclic cystine knot consisting of an interlocking arrangement of three disulfide bridges [4]. They were the first discovered family of gene-expressed CPs and remain the most populated family among all depositions in the world-wide repository of CPs called CyBase [5], which as of January 2013 includes 633 cyclic proteins from 86 species. Cyclotides are among the most structurally stable proteins and are implicated in a diverse range of functions, from plant defense [3,6] to anti-HIV, antimicrobial, hemolytic, and uterotonic capabilities [7]. They also have strong therapeutic potential and are being actively pursued as peptide-based drug leads, molecular probes, diagnostic agents, and immunosuppressants [7–11].

Besides cyclotides, two other families of CPs are trypsin inhibitors and bacteriocins. Cyclotides and trypsin inhibitors share the cystine knot motif. By contrast, bacteriocins are larger than cyclotides and trypsin inhibitors and do not contain a cystine knot. Bacteriocins exhibit various inhibitory functions, mainly against bacteria, such as inhibition of cell-wall synthesis and RNase or DNase activity [12]. Importantly, CPs can be synthetically synthesized [13] and efforts are being made to lower the corresponding production costs [14]. The abovementioned characteristics make CPs particularly desirable as potential therapeutic agents [15,16].

Recent studies show that CPs are more common in the plant kingdom than was previously thought [5], including reports which suggest that cyclotides might include thousands of members [17]. The CyBase repository is undergoing continuing growth and it is expected that it will continue growing at a substantial pace [18]. Moreover, the biosynthetic mechanism of cyclization remains uncertain, and thus information on mechanisms currently cannot be used to indicate which species, and to what degree, produce cyclic proteins. These

[☆] This article is part of a Special Issue entitled: Computational Proteomics, Systems Biology & Clinical Implications. Guest Editor: Yudong Cai.

* Corresponding author. Tel.: +1 780 492 5488; fax: +1 780 492 1811.
E-mail address: lkurgan@ece.ualberta.ca (L. Kurgan).

1570-9639/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.bbapap.2013.05.002

considerations provided the motivation for the current study, in which we design an accurate and fast in-silico method to predict whether a given protein chain is cyclic. Most importantly, this method is used to predict and characterize putative CPs on a proteomic scale across hundreds of eukaryotic, bacterial and archaea proteomes. Similar computational studies were recently carried out to characterize various functional classes of proteins, e.g., for disordered proteins [19,20], caspases [21], and zinc proteins [22].

2. Materials and methods

2.1. Datasets

We collected representative sets of data for cyclic and non-cyclic proteins. All wild-type cyclic chains, which were downloaded from CyBase [5] in July 2011, were clustered at 90% sequence similarity with CD-HIT [23] to remove redundancy; one chain from each cluster was kept. CD-HIT is a popular method (e.g., it was used to cluster UniProt to create the UniRef datasets) that implements a fast greedy incremental clustering which groups sequences into clusters that are characterized by sequence similarity above a pre-defined threshold [23]. A total of 109 cyclic chains was obtained, including 100 cyclotides, three cyclic defensins from primates, four bacteriocins, and two trypsin inhibitors. We included only CPs that exceeded 10 AAs in size, since shorter chains would be difficult to compute by our predictor, i.e., they could not be reliably represented by features that are used as its inputs. Non-cyclic proteins were extracted from the Protein Data Bank (PDB) [24], version from November 2010, using a representative subset of high-quality crystal structures that had well-separated termini. Specifically, we collected all 69,510 depositions from PDB and removed DNA and RNA strands. We excluded non-X-ray and lower quality structures to accurately calculate the distance between termini, i.e., we excluded PDB depositions with resolution > 2.00 Å and R-factor > 0.25, which is consistent with recent related studies [25,26]. Next, among the resulting 25,316 chains we removed those with unstructured (disordered) termini, i.e., all chains that lacked spatial coordinates for their first or last 10 residues. The remaining 8694 chains were processed to select those that had sufficiently separated termini. We excluded all structures for which the distance between termini was smaller than their radius; radius was defined as the distance between the center of mass (i.e., arithmetic mean of atomic coordinates) of the protein and the center of mass of the furthest from the center residue and the distance between termini was calculated between the centers of mass of the first and last residue. Finally, the 3908 structures with well-separated termini were cross-referenced against CyBase (none of these chains was found to be cyclic) and clustered at 40% sequence similarity with CD-HIT.

The resulting 683 non-cyclic proteins, together with the 109 cyclic chains, were divided into two equally-sized datasets, a TRAINING dataset that was used to design the prediction method and a TEST dataset that was utilized to perform independent (from the TRAINING proteins) evaluation of the predictive quality of the predictor. The TRAINING dataset includes 55 cyclic and 342 non-cyclic proteins and the TEST dataset has 54 cyclic and 341 non-cyclic chains; both datasets include cyclic defensins, bacteriocins, and trypsin inhibitors with 2, 2, and 1 examples, respectively, in the TEST dataset. We also collected a dataset of 23 non-redundant cyclic proteins that were deposited to CyBase after we selected cyclic chains for the TRAINING and TEST datasets, i.e., after July 2011. These proteins, which include 22 cyclotides and a small trypsin inhibitor from sunflowers, form the TEST_NEW dataset, which was used to perform additional validation of our predictor. We note that our datasets focus on the currently well-represented, i.e., having sufficient number of chains, families of wild-type CPs, including cyclotides, cyclic defensins, bacteriocins, and trypsin inhibitors. This means that the predictive model generated and evaluated with these datasets is also limited to these CP families.

Furthermore, we evaluated predictions on a representative subset of PDB. We utilized the abovementioned 8694 high-quality X-ray structures, removed duplicate chains and clustered them at 80% sequence similarity with CD-HIT. The remaining non-redundant (at 80% similarity) set of 1737 chains is named PDB80 and includes four CPs: a cyclotide, two trypsin inhibitors (including the small trypsin inhibitor from the TEST_NEW dataset), and a bacteriocin. The four datasets, including protein IDs, names, sequences, and species, are available at http://biomine.ece.ualberta.ca/CyPred/.

2.2. Evaluation measures and test protocols

Predictions of CPs were assessed by comparing against the native annotations. The evaluation was performed based on four commonly used measures:

$$\text{Sensitivity} = \text{TP}/(\text{TP}+\text{FN}) = \text{TP}/\text{N}_{\text{cyclic}}$$
$$\text{Specificity} = \text{TN}/(\text{TN}+\text{FP}) = \text{TN}/\text{N}_{\text{non-cyclic}}$$
$$\text{Accuracy} = (\text{TP}+\text{TN})/(\text{TP}+\text{TN}+\text{FP}+\text{FN}) = (\text{TP}+\text{TN})/\text{N}$$
$$\text{MCC} = (\text{TP}*\text{TN}–\text{FP}*\text{FN})/\text{sqrt}\{(\text{TP}+\text{FP})*(\text{TP}+\text{FN})*(\text{TN}+\text{FP})*(\text{TN}+\text{FN})\}$$

where TP and TN are the counts of true positives (correctly predicted cyclic proteins) and true negatives (correctly predicted non-cyclic proteins), respectively; FP and FN are the numbers of false positives (non-cyclic proteins predicted as cyclic) and false negatives (cyclic proteins predicted as being non-cyclic), respectively; and N, N_{cyclic} and $\text{N}_{\text{non-cyclic}}$ are the total counts of all, cyclic and non-cyclic chains in a given dataset, respectively. The Matthews correlation coefficient (MCC) was suggested to be used to assess predictions in cases where the numbers of positive (cyclic) and negative (non-cyclic) samples are substantially different [27], which is true for our datasets. The MCC values range between −1 and 1, with 0 denoting random prediction and higher absolute values denoting more accurate predictions. Higher values of the other three measures indicate better quality of predictions.

We designed a predictor of cyclic chains, i.e., we selected features that are used to encode the input protein chains and parameterized the classification models, using 5-fold cross validation on the TRAINING dataset. We split the dataset into five equal-sized subsets of protein chains and used four of these subsets to form a training fold that was utilized to compute the model and the remaining subset was used to perform the evaluation. This was repeated five times, each time choosing a different subset to be the test fold. The tests on the TEST, TEST_NEW and PDB80 datasets were based on the model that was generated on the TRAINING dataset.

2.3. Prediction model

The predictions were performed in two steps. First, the input protein chain was converted into a small set of numerical features. Next, these features were inputted into a classification model that generates the prediction. We considered a relatively simple feature-based sequence representation, which was motivated by the need to perform the predictions in a high-throughput fashion. The features included:

- amino acid (AA) composition, defined as the fraction of AA of a particular type in a given protein chain (20 features),
- hydrophobicity and charge; using hydrophobicity we divided the AAs into hydrophobic, hydrophilic or neutral and using charge we split the AAs into positive, negative, or neutral based on the categorization of AAs from [28]; we calculated the composition for each of these six sets of AAs (six features),
- frequency of certain sequence motifs in the input protein chain (six features that are explained below),
- content of secondary structures and normalized number of secondary structure segments predicted with PSI-PRED [29]; the content is defined as a fraction of residues predicted to be in the coil, strand, and helix conformations; the number of helical, coil and strand segments is divided by the total number of all secondary structure segments (six features).

The use of the AA composition and hydrophobicity and charge-defined groups is supported by the fact that CPs differ in their composition from non-cyclic proteins, as shown in Fig. 1. We observe that cyclotides and trypsin inhibitors, which have the characteristic cystine knot, are substantially enriched in Cys compared to the non-cyclic proteins. They are also enriched in Gly and Pro and depleted in Met, Glu, Gln, Ala, Leu, and Asp. The bacteriocins are enriched in Ile, Thr, and Ala, and depleted in Cys, Pro, Tyr, Asn, Arg, Glu, and Asp when compared to the non-CPs. These differences provided useful predictive inputs.

We hypothesized that cyclic chains might have a particular arrangement of secondary structures that would allow their differentiation from other proteins. PSI-PRED was selected due to its relatively strong predictive performance [30] and the definitions of features were motivated by their successful prior use to predict major classes of protein folds [31,32].

Moreover, motivated by the fact that cyclic proteins have several characteristic sequence motifs, we also designed additional features that detect these motifs in the input chain:

- The presence of cystine knot motifs (three binary features). Cyclotides and trypsin inhibitors have the cystine knot that consists of six Cys residues connected by three disulfides bridges. Cyclotides are divided in two structural subfamilies which differ by the presence of a cis-proline in loop five (Möbius vs. bracelet). The Cys residues are spaced in a specific way in the chain, depending on the structural subfamily [33]:

  CxxxCxxxxCxxxxxxCxCxxxxC is the conserved motif in bracelet type cyclotides

  CxxxCxxxxCxxxxCxCxxxxC for the Möbius type cyclotides

  We used three binary values (present vs. absent) to indicate whether one of the two, or either, motif is present in the input chain.
- The deviation from the cystine knot motif (one real-valued feature). This feature quantifies how well a given chain adheres to the CxxxCxxxxCxxxxxxCxCxxxxC motif. This is calculated by counting the number of AAs between two consecutive Cys residues in a given chain and subtracting from the actual number of residues in between the corresponding two Cysteines in the motif, e.g., for a CtaetC motif in the input chain, this would be $4 - 3 = 1$). The absolute values of these differences for each pair of Cys residues are calculated and summed.
- The presence of the CGES(T)C motif (one binary feature). This is a conserved motif in the loop 1 (between Cys I and II) in the bracelet and Möbius types of cyclotides. We use a value of 1 if CGESC or CGETC fragment occurs in the input chain; 0 otherwise.
- The composition of Cys in a sequence window (1 real-valued feature). We compute the maximal number of Cys residues in a sliding window of 30 AAs in the input chain.

A total of 38 features was used, including the 20-dimensional AA composition, six charge/hydrophobicity residue groups, six secondary structure-based features, and the six motif-based features. We evaluated four classification algorithms that generate predictive models. They included a classical logistic regression and three modern classifiers, which are listed among the top 10 data mining algorithms [34], including Naïve Bayes, Support Vector Machine (SVM), and C4.5 decision tree. We compared the predictive performance of these four classifiers using several combinations of feature sets to select the setup that provides the strongest predictive performance. We tested the use of 20-dimensional AA composition, AA composition combined with motifs (26 features), these 26 features combined with either secondary structure-based features (32 features) or hydrophobicity/charge-based features (32 features, and all 32 features. The results based on the 5-fold cross validation on the TRAINING dataset are shown in Table 1. The SVM model uses a popular Radial Basis Function kernel and was parameterized utilizing grid search with complexity constant $C = 2^i$, $i = -1, 0, \ldots 10$, and $gamma = 2^k$, $k = -5, -4, \ldots, 5$ based on maximization of MCC with the 5-fold cross validation on the TRAINING dataset.

As expected, the SVM model provides the most accurate predictions, i.e., the highest values of MCC and accuracy. This is because SVM utilizes an optimized non-linear model compared to the other three considered classifiers that use simpler linear models. Interestingly, the results also reveal that use of the motif-based features slightly improves the predictive performance of the SVM model when compared with the use of just the AA composition. In particular, this set of 26 features results in predictions with higher specificity while maintaining the same sensitivity, which means that the number of false positives (non-cyclic proteins predicted as cyclic) was reduced. The same trend is also true for the second-best logistic regression model. Addition of the hydrophobicity/charge-based and the secondary structure-based features does not provide further improvements. Although SVM that uses these additional features improves sensitivity to 100%, this is coupled with a drop in specificity such that the overall accuracy and MCC are slightly lower than when using the 26 features. We note that CPs assume a relatively wide range of secondary structure arrangements, from hairpins to all-helical structures, which is likely why the use of the secondary structure did not help. Moreover, the prediction of the secondary structures is time consuming due to the generation of the position specific scoring matrices (PSSMs) by PSI-PRED, which constitutes a drawback associated with a potential inclusion

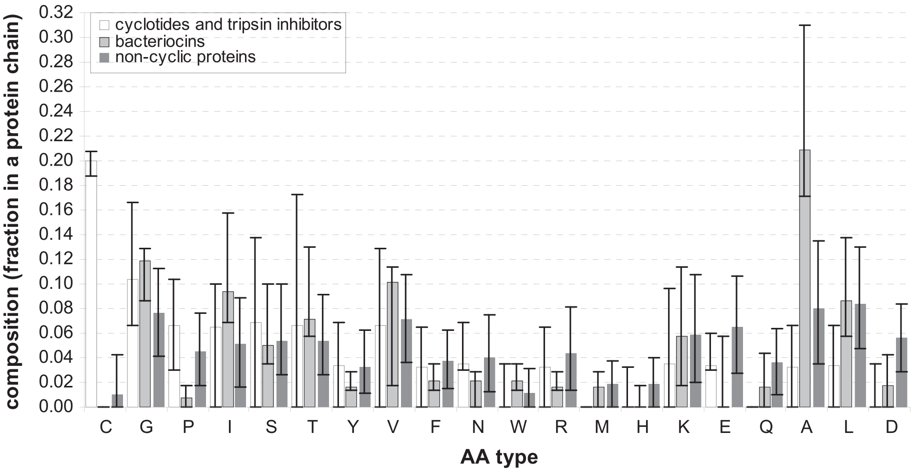

Fig. 1. Comparison of composition of amino acid (AA) types between cyclic and non-cyclic proteins using chains from TRAINING and TEST datasets. Bars represent the 50th centile (median) of the compositions values across all chains, and the error bars show the 10th centile and 90th centile. AA types are sorted based on the values of the difference between the medians for all cyclic and non-cyclic chains, from AA types enriched in cyclic proteins on the left to those enriched in non-cyclic proteins on the right.

Table 1
Comparison of predictive performance of the four considered classifiers based on the 5-fold cross validation on the TRAINING dataset. The classifiers are evaluated when using the 20-dimensional AA composition in combination with 6 sequence motif-based features, 6 features based on predicted secondary structure, and 6 features based on hydrophobicity and charge-based residue groups, as their inputs.

Input features	Prediction model	MCC	Accuracy	Sensitivity	Specificity
38 features	SVM	0.97	99.2	100.0	99.1
(AA composition, motifs, secondary structure,	Logistic Regression	0.86	96.7	93.8	97.1
hydrophobicity and charge)	C4.5	0.94	98.5	94.5	99.1
	Naïve Bayes	0.92	98.0	91.2	99.1
32 features	SVM	0.97	99.2	100.0	99.1
(AA composition, motifs, and secondary structure)	Logistic Regression	0.92	98.2	98.0	98.3
	C4.5	0.94	98.5	94.5	99.1
	Naïve Bayes	0.92	98.0	91.2	99.1
32 features	SVM	0.98	99.5	98.2	99.7
(AA composition, motifs, and hydrophobicity and charge)	Logistic Regression	0.96	99.0	96.4	99.4
	C4.5	0.94	98.5	94.5	99.1
	Naïve Bayes	0.94	98.5	93.0	99.4
26 features	SVM	0.98	99.5	98.2	99.7
(AA composition and motifs)	Logistic Regression	0.97	99.2	94.5	100.0
	C4.5	0.94	98.5	94.5	99.1
	Naïve Bayes	0.93	98.2	96.4	98.5
20 features	SVM	0.97	99.2	98.2	99.4
(AA composition)	Logistic Regression	0.94	98.5	94.5	99.1
	C4.5	0.97	99.2	96.4	99.7
	Naïve Bayes	0.93	98.2	96.4	98.5

of the corresponding features. We also note that the predictive quality obtained with the C4.5 classifier slightly drops with the inclusion of additional features, which could be a result of an overfitting. Consequently, the SVM classifier with the 26 features was used to implement the proposed predictor of cyclic proteins, which is named CyPred.

The architecture of CyPred is shown in Fig. 2. The CyPred method outputs real-valued confidence scores, where larger positive/negative value indicates higher propensity to be cyclic/non-cyclic. A web server that implements CyPred is publicly available at http://biomine.ece.ualberta.ca/CyPred/.

3. Results and discussion

3.1. Evaluation of predictive quality and runtime on test datasets

The predictive quality of CyPred was compared with currently available approaches to identify cyclic proteins, which include sequence alignment methods. A given test sequence was aligned against all sequences (both cyclic and non-cyclic) from the TRAINING dataset and the label of the most similar training sequence was assigned as the prediction. This allowed for a side-by-side comparison with CyPred that also uses the TRAINING dataset to build the prediction model. The similarity was measured by the number of aligned matching AAs divided by the length of the test or training chain, whichever was shorter, and we selected the training chain with the largest score to transfer the label. We used two popular types of alignment methods:

- The BLAST algorithm [35] with default parameters; in cases when no alignments were returned we classified the corresponding chain as non-cyclic;
- Pairwise alignment with the Smith-Waterman-Gotoh algorithm [36,37]; we aligned a given chain against each chains in the TRAINING dataset.

We also compared the full implementation of CyPred with a version that utilizes only the 20-dimensional AA composition. The results on the TEST dataset are summarized in Table 2. CyPred obtains 100% sensitivity and the highest specificity, which equals to the specificity of BLAST. Our method correctly predicted all 54 cyclic proteins and generated five false positives (non-cyclic chains predicted as cyclic). To compare, SVM with 20 features, BLAST, and pairwise alignment produced 8, 5, and 32 false positives, respectively. Importantly, CyPred correctly predicted all native cyclic proteins, including two cyclic bacteriocins and a trypsin inhibitor. This demonstrates that our method is capable of finding CPs from various families, besides the most populated (cyclotides). On the other hand, BLAST incorrectly identified one of the bacteriocins and the trypsin inhibitor as non-cyclic proteins, while pairwise alignment incorrectly predicted two bacteriocins as being non-CPs. Furthermore, the use of the 6 motif-based features improved predictions of CyPred when compared with the SVM that uses only the 20-dimensional composition. Consistent with the results on the TRAINING dataset, see Table 1, we observed an increase in specificity, which means that fewer false positives were generated by CyPred, thanks to the use of these motifs.

We predicted the 23 nonredundant CPs from the TEST_NEW dataset using CyPred. All these chains were correctly identified as cyclic with high confidence scores; the lowest score was 0.55 and 21 out of 23 chains were predicted with scores above 0.9. The small trypsin inhibitor that was included in this set obtained score of 1.0.

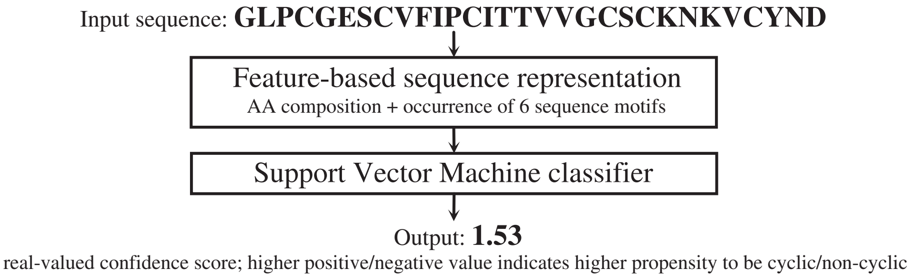
Input sequence: GLPCGESCVFIPCITTVVGCSCKNKVCYND
Feature-based sequence representation
AA composition + occurrence of 6 sequence motifs
Support Vector Machine classifier
Output: 1.53
real-valued confidence score; higher positive/negative value indicates higher propensity to be cyclic/non-cyclic

Fig. 2. Architecture of the CyPred method.

Table 2
Comparison of predictive performance of CyPred, SVM-model that uses AA composition, BLAST and pairwise alignment on the TEST dataset. The methods are sorted by their MCC scores.

Prediction model	MCC	Accuracy	Sensitivity	Specificity
CyPred	0.95	98.7	100	98.5
BLAST	0.93	98.2	96.3	98.5
SVM with AA composition (20 features)	0.92	98.0	100	97.6
Pairwise alignment	0.73	91.4	96.3	90.6

We also assessed the runtime of the two most promising, according to the predictive quality, approaches: CyPred and BLAST. The runtime was evaluated using a modern desktop computer on the TEST dataset. Although the values depended on the hardware used, we concentrated on the relative differences in the runtime between CyPred and BLAST, which should be hardware-independent. Fig. 3 shows that BLAST is characterized by more than an order of magnitude longer runtime, which slightly increases with the length of the input protein chain. The ratio between the BLAST and CyPred runtime ranged between 20:1 for short chains and 27:1 for long chains. The runtime of CyPred varied between 2 and 6 ms per protein, which means that it can be used to perform high-throughput predictions.

3.2. Predictions on the PDB80 dataset

CyPred, BLAST, and pairwise alignment were used to predict CPs in the PDB80 dataset, which is a representative subset of 1737 PDB chains. We compared predictive performance of these three approaches based on two factors: predictive quality measured with MCC, accuracy, sensitivity, and specificity; and distribution of distances between termini in the corresponding PDB structures for the chains predicted to be cyclic; see Table 3.

We observed that the predictive performance of all considered approaches was lower on this dataset compared to the TEST dataset. This could be explained by the fact that proteins in the TRAINING dataset share higher sequence similarity with those in the TEST dataset, in contrast to their similarity with the chains in the PDB80 set. Importantly, CyPred consistently (over both TEST and PDB80 datasets) provided improved predictive performance. In the PDB80 dataset, our method correctly predicted three of four cyclic proteins (75% sensitivity), which included a cyclotide, a small trypsin inhibitor, and a bacteriocin. This further (similarly as with the TEST dataset) demonstrates that CyPred successfully predicts different families of CPs. BLAST and pairwise alignment obtain the same sensitivity, but their specificity is lower. To compare, CyPred, BLAST and pairwise alignment generated 8, 10, and 60 false positives, respectively. Moreover, the chains predicted by CyPred as cyclic have substantially smaller distances between their termini, i.e., the median distance for CyPred is 6.5 Å, while the median distances for putative CPs produced by BLAST and pairwise alignment are 11.9 Å and 16 Å, respectively. The corresponding median distance across all proteins in the PDB80 dataset equals 25.7 Å. Distributions of these distances are visualized in Fig. 4A and they demonstrate that CyPred generates higher quality predictions. Fig. 4B shows that the incorrect predictions generated by CyPred are mostly for chains with a low distance between termini, while BLAST and pairwise alignment generate more errors for chains characterized by larger inter-terminal distances. The observation that CyPred generates false positives characterized by short distances between termini suggests that our predictor can be also used to find such chains. These chains could be artificially cyclized, which would assist with finding proteins of interest (e.g., certain enzymes or toxins) that are cyclizable.

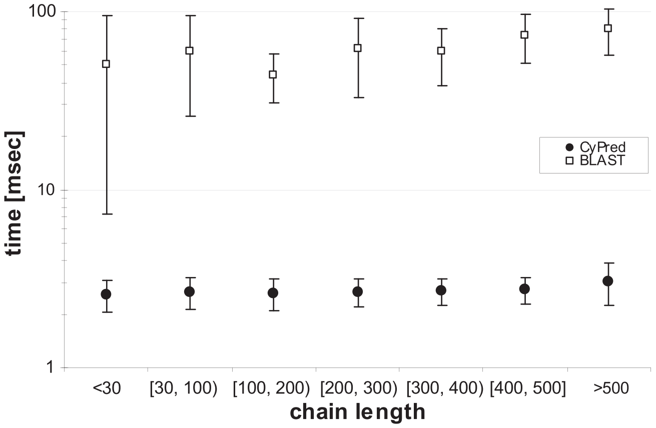

Fig. 3. Runtime of CyPred and BLAST calculated based on predictions on the TEST dataset. Proteins in the TEST dataset were divided by their chain length (*x*-axis) and the runtime (*y*-axis in milliseconds shown using logarithmic scale) is shown as average values (markers) over the chains in a given size interval with the corresponding standard deviations (error bars).

The cyclic proteins predicted by CyPred include three true positives (TP; correctly predicted CPs) and eight false positives (FPs), which corresponds to a precision = TP/(TP + FP) = 27%. We use this level of precision to estimate the number of CPs when performing predictions on the whole proteomes. We also observed that predictions with high confidence values are more likely to include native CPs and we used cut-off value of 0.9 to indicate cyclic chains predicted with high quality. Fig. 4A and B shows that all chains predicted by CyPred with confidence scores > 0.9 have their distances between termini below 10 Å and that they include only three incorrect predictions.

3.3. Cyclic proteins in the three domains of life

CyPred, which offers favorable predictive performance and faster predictions compared to the alignment, was used to predict CPs in 640 complete proteomes from the three domains of life. The corresponding 5,700,468 proteins were collected from release 2011_08 of UniProt [38]. The proteomes were assigned to their taxonomic lineage based on NCBI [39]. We also collected two eukaryotic proteomes, from *Violaceae* and *Rubiaceae* plants, from UniProt as they are from families with the largest known populations of cyclotides. The putative CPs predicted by CyPred were filtered to remove chains that were deleted in the UniProt since release 2011_08 (as of March 2013 when we performed filtration) and chains annotated as predicted with a caution, which may indicate serious problems like frameshifts, errors in initiation or stop codons, error in translation, etc. This resulted in removal of 515 chains out of the original set of 4014 putative CPs; the remaining putative CPs are summarized in Table 4.

Using the precision of 27%, which was estimated from the PDB80 dataset, CyPred found total of 3499 cyclic proteins from among 5.7+ million chains in the considered 642 proteomes. We note that this set of putative CPs is potentially incomplete since CyPred is designed using data that includes only certain families of CPs, such as cyclotides, cyclic defensins, bacteriocins, and trypsin inhibitors; this means that other families of CPs are likely excluded from the results. Although CPs are relatively rare, i.e., only about 0.06% of chains are predicted to be cyclic, they were found in the majority of proteomes in the three domains of life. More precisely, between 89 and 98% of proteomes, depending on the domain of life, have at least one predicted CP; and between 45 and 56% of proteomes include at least one CP predicted with high confidence (with the confidence score > 0.9, which indicates high predictive quality). However, only a small fraction of proteomes have larger counts of CPs, i.e., only between 7% (for archaea proteomes) and 16% (for eukaryotic proteomes) of proteomes have over 10 cyclic proteins. There are no proteomes in archaea with more than 10 CPs predicted with high confidence, while 2% and 11% of proteomes in bacteria and eukaryota, respectively, have at least 10 CPs that were identified with the high confidence.

Table 3
Comparison of the predictive performance of CyPred, BLAST and pairwise alignment on the PDB80 dataset. The methods are sorted by their MCC scores. The last row shows the median and distribution of distances between termini across all structures from the PDB80 dataset.

Prediction model/statistics using the PDB80 dataset	Median distance between termini	% chains with distance between termini			MCC	Accuracy	Sensitivity	Specificity
		<20 A	<10 A	<5 A				
CyPred	6.47	91.0%	72.7%	45.5%	0.45	99.48	75.00	99.54
BLAST	11.87	62.0%	38.5%	23.1%	0.41	99.37	75.00	99.42
Pairwise alignment	16.03	76.0%	19.0%	7.9%	0.18	96.49	75.00	96.54
Entire PDB	25.65	35.0%	9.9%	1.2%				

Fig. 5 provides a more detailed breakdown of the number of CPs per proteome. The majority of proteomes have very few cyclic proteins, i.e., three or fewer for archaea proteomes, and two or fewer for bacterial and eukaryotic proteomes. However, 10% of proteomes have at least 8, 11, or 14 CPs for archaea, bacterial, or eukaryotic proteomes, respectively. We note the relatively large numbers of CPs that were predicted with high confidence in some eukaryotes and bacteria. Specifically, 6% of the considered eukaryotic proteomes (seven proteomes) and 1% of bacterial proteomes (five proteomes) have at least 20 such putative cyclic chains. Fig. 6 summarizes the similarity of all 525 predicted CPs from these 12 CP-enriched proteomes to the native cyclotides, bacteriocins, and trypsin inhibitors. The markers, which represent individual putative CPs, are grouped by proteomes (shown on *x*-axis), and their values (*y*-axis) correspond to a similarity to the closest CP family. The similarity was quantified with Euclidian distance between the average AA composition of chains in a given CP family (using data from the TRAINING and TEST datasets) and the AA composition of a given putative CP. The predicted CPs that have similarity lower than the cut-off of 0.265 are assigned to the corresponding CP family; the cut-off value corresponds to an average distance between native CPs that belong to different CP families. The results show that most of the CPs predicted for the five bacteria (shown on the left) are similar to bacteriocins,

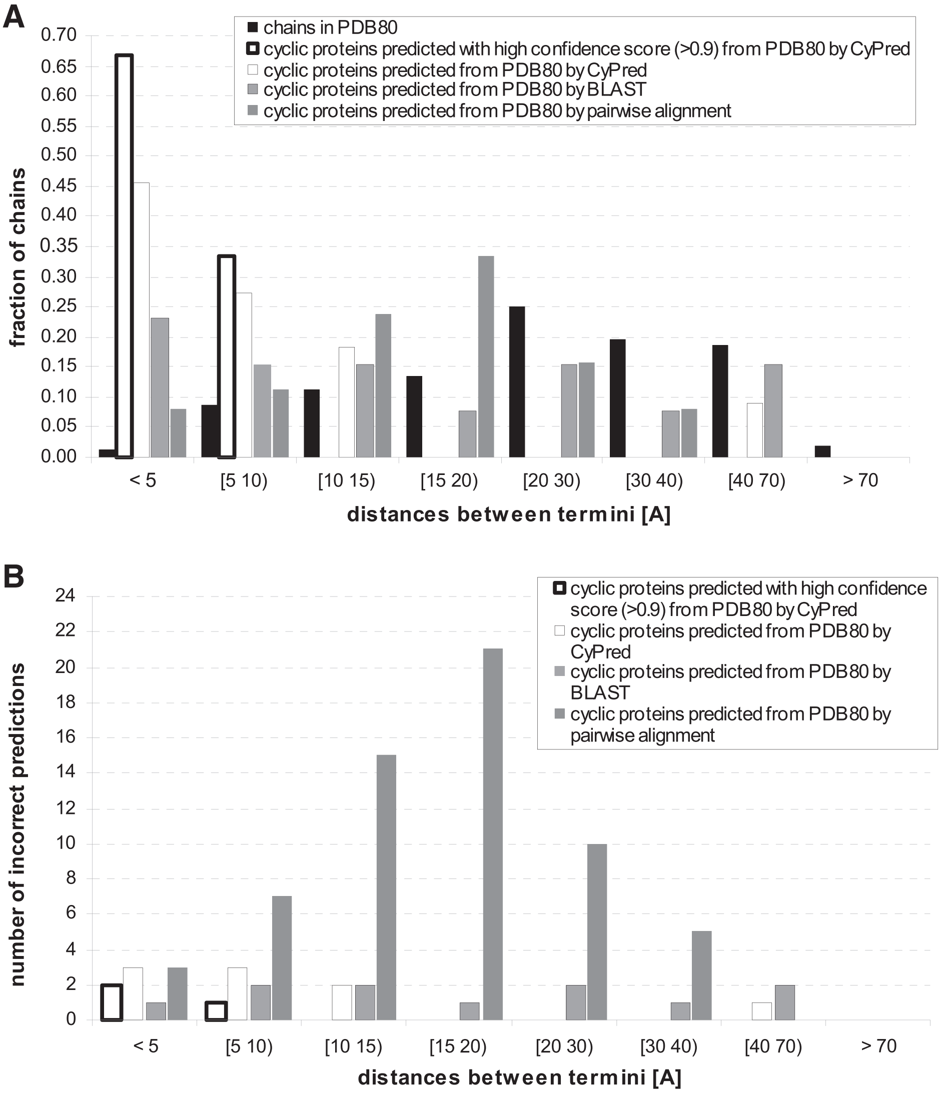

Fig. 4. Distribution of distances between termini for chains in PDB dataset and chains predicted as cyclic from the PDB80 dataset (panel A) and number of incorrect predictions (panel B) by CyPred, BLAST and pairwise alignment. The distances were binned into intervals shown on the *x*-axis.

Table 4
Summary of results from the prediction of cyclic proteins by CyPred on the 640 complete proteomes from the 2011_08 release of UniProt and proteomes of *Violaceae* and *Rubiaceae* plants. The first two rows summarize the considered proteins; the third row shows the results over all proteomes in a given domain of life; the fourth row summarizes results per proteomes; and the last row summarizes the predictions with high confidence (with scores > 0.9) per proteomes.

		Archaea	Bacteria	Eukaryota
Number of considered proteomes		59	471	112
Total count of considered proteins		216,370	3,627,676	1,856,422
Predicted cyclic proteins	Number	243	2537	719
	% of all proteins	0.11%	0.07%	0.04%
Predicted cyclic proteins per proteomes	Number of proteomes with at least one cyclic protein	58	463	100
	% of proteomes with at least one cyclic protein	98.3%	98.3%	89.3%
	% of proteomes with >5 cyclic proteins	30.5%	23.4%	27.7%
	% of proteomes with >10 cyclic proteins	6.8%	11.0%	16.1%
cyclic proteins predicted with high confidence (score > 0.9) per proteomes	Number of proteomes with at least one cyclic protein	33	237	51
	% of proteomes with at least one cyclic protein	55.9%	50.3%	45.5%
	% of proteomes with >5 cyclic proteins	3.4%	6.2%	15.2%
	% of proteomes with >10 cyclic proteins	0.0%	2.3%	10.7%

while majority of the putative CPs in eukaryotes are similar to cyclotides. There are also a few CPs that are similar to trypsin inhibitors for several bacteria and eukaryotas. A total of 132 out of the 525 putative CPs have a distance above the cut-off (shown using crosses in Fig. 6), which suggests that they are dissimilar to the three major families of CPs. The largest number of these CPs, 42, was found in a *Nematostella*, which is a sea anemone.

We also analyzed the distribution of Cys content among the predicted CPs (Fig. 7). We observed that cysteines are substantially enriched in eukaryotes and this follows the trend for native CPs in the TRAINING and TEST datasets. Our native cyclic proteins mostly include cyclotides, which have the cystine knot that, in turn, results in the enrichment in cysteines. Compared to 392 chains in eukaryotic proteomes that have at least six Cys and were predicted to be cyclic, only one putative CP in archaea and 31 in bacterial proteomes have at least six Cys residues. This suggests that only eukaryotic proteomes contain a large fraction of cyclotides.

Fig. 8 presents the relations between the fractions of predicted CPs (*y*-axis), the fraction of CPs predicted with high confidence (*x*-axis) and the corresponding proteome size (size of markers) across all proteomes that have cyclic chains. The proteomes/markers are distributed on an approximately diagonal line, which means that the number of predictions with high confidence is proportional to the overall number of predictions, with the Pearson correlation coefficient of 0.97. The proteome size and the number of predicted cyclic chains are weakly correlated, with the exception of the archaea where there is no correlation, i.e., the Pearson correlation coefficients are 0.09, 0.31, and 0.33 for archaea, bacterial, and eukaryotic proteomes, respectively. The differences in the fractions of predicted CPs between proteomes can be as large as three orders of magnitude. The archaea proteomes have overall relatively high rates of predicted CPs, which is coupled with their relatively compact proteome sizes that vary between about 500 and 5000 proteins. However, there are some fairly large eukaryotic proteomes, with over 20000 proteins, that also have large fractions of CPs. Among eukaryotic proteomes, animals have higher ratios of CPs predicted with high confidence compared with fungi while plants have the largest spread of the ratios, ranging from some plants that have no putative CPs (see caption for Fig. 8) to the largest ratios of cyclic proteins for proteomes of *Violaceae* and *Rubiaceae* (see inset in Fig. 8). These two plant families are known to have a large number of cyclic proteins, and our analysis in Fig. 6 confirms this. They are located closest the top right corner in Fig. 8, which indicates that they have the highest fraction of predicted CPs among all considered proteomes.

We also investigated the relationship between the fraction of CPs predicted with high confidence per proteome and evolutionary speed (see Fig. 9). The evolutionary speed corresponds to the branch length in the evolutionary tree taken from [40], where larger values

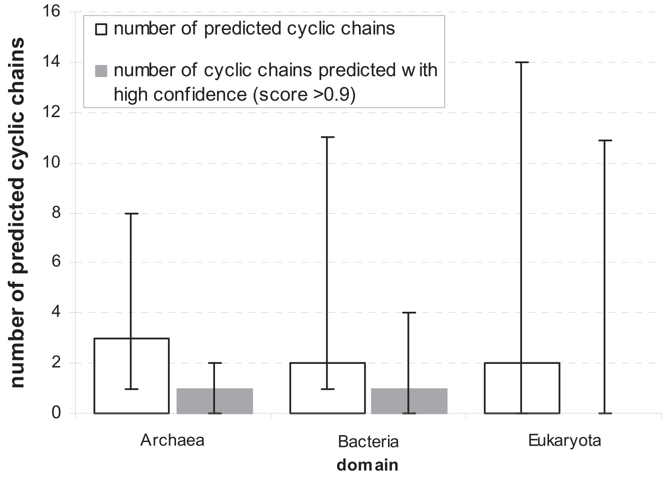

Fig. 5. Comparison of the number of predicted cyclic proteins per proteomes between the three domains of life. Bars represent the 50th centile (median) number of cyclic proteins per proteomes in a given domain and the error bars show the 10th centile and 90th centile.

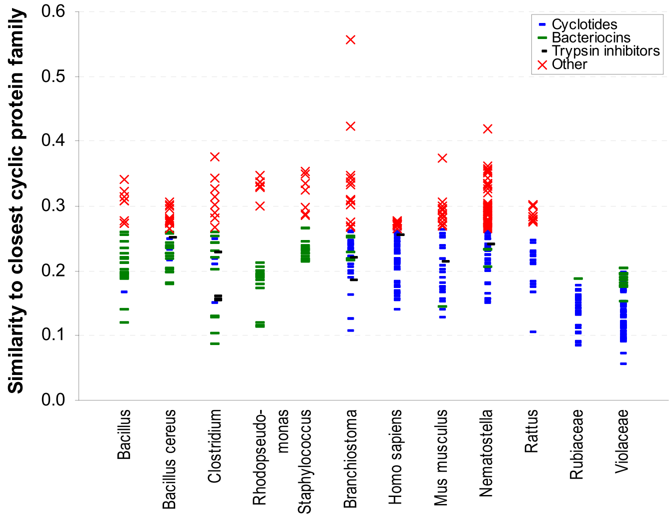

Fig. 6. Similarity of putative CPs from 12 CP-enriched proteomes to the native cyclotides, bacteriocins, and trypsin inhibitors. Each marker represents an individual putative CPs. The results are grouped by proteomes (shown on *x*-axis); five left-most proteomes are from bacteria and the remaining seven proteomes are from eukaryota. The similarity (shown on *y*-axis) was quantified with Euclidian distance between the average AA composition of chains in a given CP family (using data from the TRAINING and TEST datasets) and the AA composition of a given putative CP. Each putative CP that has similarity lower than a cut-off = 0.265 is assigned to the corresponding color-coded CP family; otherwise it is assumed to be dissimilar and is shown using red crosses. The cut-off value equals to an average distance between native CPs that belong to different CP families.

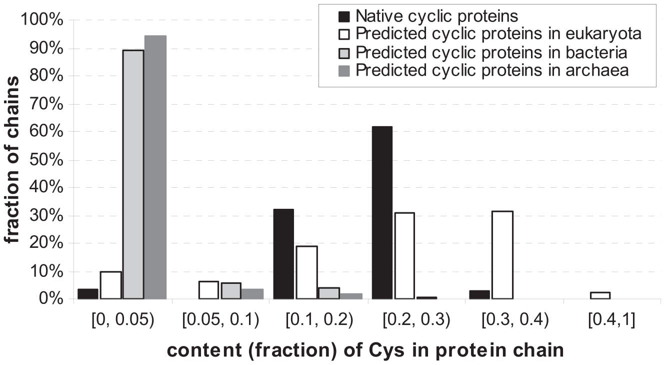

Fig. 7. Comparison of distribution of the Cys composition (fraction of Cys in a given chain) between native cyclic proteins from TRAINING and TEST datasets and cyclic proteins predicted with high-confidence (score > 0.9) by CyPred in eukaryotic, bacterial, and archaea proteomes.

denote proteomes that have a faster evolutionary rate. We were able to map 57 bacterial, 11 eukaryotic, and 1 archaea (not shown due to the small sample size) proteomes into the organisms that were included in [40]. Data for bacterial and eukaryotic proteomes show a consistent trend, with a negative correlation between the fraction of CPs and evolutionary speed, i.e., the Pearson correlation coefficients equal −0.45 and −0.44 for bacterial and eukaryotic proteomes, respectively. Although the magnitude of the correlation coefficient is relatively modest, the overall trends across the two domains are similar and reveal that proteomes that are enriched in CPs evolve slower than proteomes that have a few or no cyclic chains.

4. Conclusions

We designed and empirically tested a novel model, called CyPred, that predicts whether a protein chain is cyclic. The prediction model focuses on the four currently well-populated families of CPs (cyclotides, cyclic defensins, circular bacteriocins, and trypsin inhibitors). Prediction of other families of CPs will be addressed in the future as more annotated data becomes available.

Empirical results on TRAINING and TEST datasets showed that CyPred achieves MCC = 0.95 and sensitivity = 100%, and outperforms other methods, including sequence alignment, logistic regression, decision trees, and Naïve Bayes. CyPred correctly identified all 23 cyclic proteins that were deposited into CyBase after the method was developed. Further tests on a more challenging and larger set of over 1500 non-redundant and high-resolution proteins collected from the PDB demonstrated that CyPred outperforms commonly used alignment methods, including BLAST and pairwise alignment. Our method obtains a MCC of 0.45 compared to 0.41 and 0.18 obtained by BLAST and pairwise alignment, respectively; the three approaches have the same sensitivity of 75%. Moreover, CyPred's predictions are characterized by shorter runtime than BLAST predictions and substantially lower distances between termini of the predicted cyclic proteins. The median distance for CyPred predictions was 6.5 Å, compared to 11.9 Å and 16.0 Å for BLAST and pairwise alignment, respectively.

Although CyPred, like other predictors including the commonly used sequence alignments, may sometimes provide incorrect predictions, it generates predictions very quickly and offers relatively good predictive quality. The above features, combined with the pressing need to explore cyclic proteins on a large scale, motivated us to use CyPred to estimate the abundance of cyclic proteins (in particular the four well-represented families of CPs) in 642 fully sequenced proteomes in the three domains

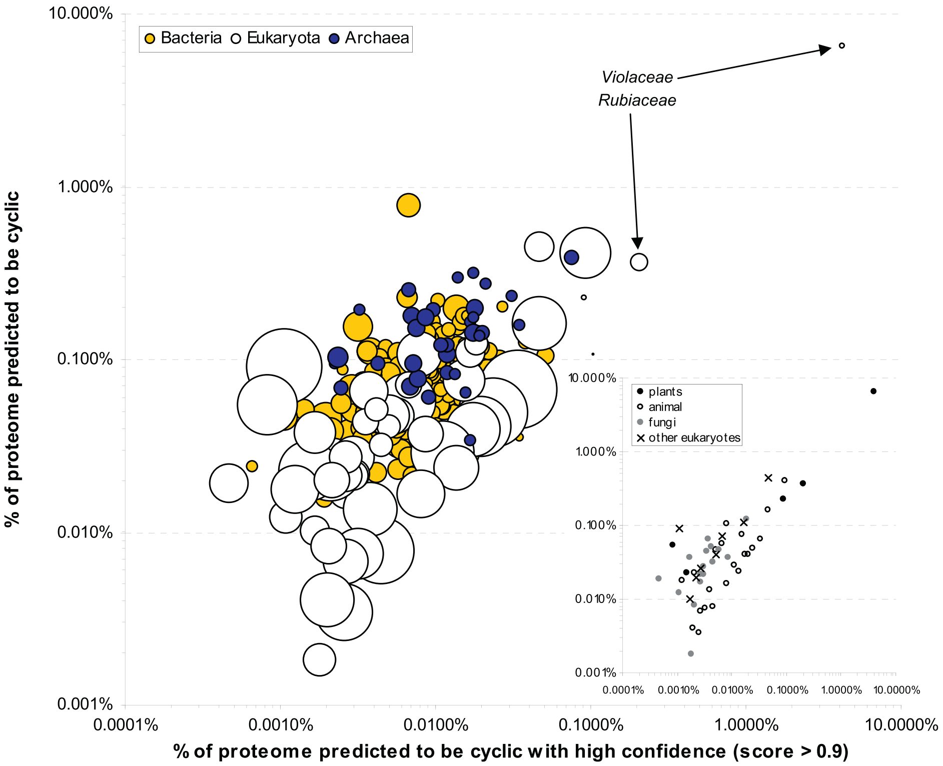

Fig. 8. Relationship between the fractions of predicted cyclic proteins (y-axis), the fractions of cyclic proteins predicted with high confidence (score > 0.9; x-axis), and the proteome size (size of markers) in proteomes across the three domains of life (color of markers). The proteome sizes range from about 55000 (for some eukaryotes) to a few of hundred of proteins (in some bacteria and archaea) and the sizes of markers are proportional to the proteome sizes. Both axes are in logarithmic scale to enhance visualization of the differences; consequently, 61 eukaryotic, 234 bacterial, and 26 archaea proteomes without proteins predicted with high confidence had to be excluded from this graph. The inset in bottom-right corner shows relation across various phyla/kingdoms in eukaryotes; 2 plants, 4 animals, 40 fungi, and 14 other eukaryotic proteomes without proteins predicted with score > 0.9 were excluded.

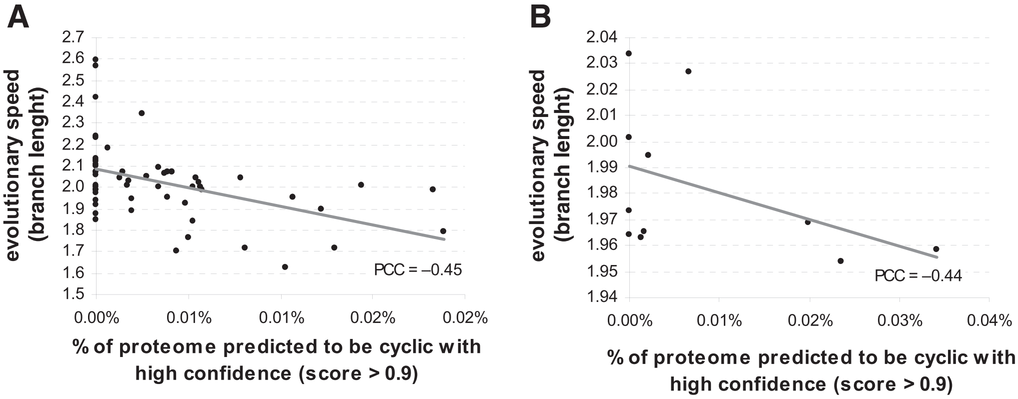

Fig. 9. Relationship between fractions of cyclic proteins predicted with high confidence (score > 0.9; *x*-axis) and evolutionary speed (*y*-axis) for proteomes in bacteria (panel A) and in eukaryote (panel B). Solid lines show linear fit together with the corresponding value of the Pearson correlation coefficient (PCC).

of life. Our analysis suggested that there are about 3500 putative CPs among 5.7+ million chains collected from these proteomes. The majority of the CPs are in bacteria and eukaryotes, and we found 74 proteomes that are predicted to have 10 or more cyclic proteins. The differences in the estimated fraction of CPs per proteome are as large as three orders of magnitude. We found relatively large numbers of CPs that were predicted with high confidence in eukaryotes and bacteria, with seven and five proteomes that have 20 or more such cyclic proteins, respectively; these putative CPs probably include a large number of cyclotides (in eukaryotes) and bacteriocins (in bacteria). Animal proteomes have higher ratios of CPs predicted with high confidence compared to fungi while the highest rates of putative CPs, which likely primarily include cyclotides, were found in certain plants. We also found that proteomes with higher ratios of CPs evolve at a slower pace than proteomes that have fewer cyclic chains. Given that only 600+ cyclic chains are currently characterized and deposited in CyBase, we conclude that further research is required to fully reveal the scope and appreciate the potential of this protein family.

A web server and a standalone implementation of CyPred together with the putative cyclic proteins extracted from the 642 proteomes are publicly available at http://biomine.ece.ualberta.ca/CyPred/.

Acknowledgements

We express sincere thanks to the organizers of the 2010/11 Sanofi BioGENEius Challenge Canada where initial results of this study were presented, and to Zhenling Peng who helped with the collection of data for the evolutionary pace analysis. MJM was supported by the University of Alberta Dissertation Scholarship. DJC is grateful for the support of a National Health and Medical Research Council (Australia) Fellowship and grant (APP1049928 and APP1047857).

References

[1] M. Trabi, D.J. Craik, Circular proteins-no end in sight, Trends Biochem. Sci. 27 (3) (2002) 132–138.

[2] L. Cascales, D.J. Craik, Naturally occurring circular proteins: distribution, biosynthesis and evolution, Org. Biomol. Chem. 8 (22) (2010) 5035–5047.

[3] D.J. Craik, Circling the enemy: cyclic proteins in plant defence, Trends Plant Sci. 14 (6) (2009) 328–335.

[4] D.J. Craik, N.L. Daly, T. Bond, C. Waine, Plant cyclotides: a unique family of cyclic and knotted proteins that defines the cyclic cystine knot structural motif, J. Mol. Biol. 294 (5) (1999) 1327–1336.

[5] C.K. Wang, Q. Kaas, L. Chiche, D.J. Craik, CyBase: a database of cyclic protein sequences and structures, with applications in protein discovery and engineering, Nucleic Acids Res. (2008) D206–D210.

[6] D.J. Craik, Host-defense activities of cyclotides, Toxins 4 (2) (2012) 139–156.

[7] A. Gould, Y. Ji, T.L. Aboye, J.A. Camarero, Cyclotides, a novel ultrastable polypeptide scaffold for drug discovery, Curr. Pharm. Des. 17 (38) (2011) 4294–4307.

[8] C. Gründemann, J. Koehbach, R. Huber, C.W. Gruber, Do plant cyclotides have potential as immunosuppressant peptides? J. Nat. Prod. 75 (2) (2012) 167–174.

[9] D.J. Craik, J.E. Swedberg, J.S. Mylne, M. Cemazar, Cyclotides as a basis for drug design, Expert Opin. Drug Discov. 7 (3) (2012) 179–194.

[10] A.B. Smith, N.L. Daly, D.J. Craik, Cyclotides: a patent review, Expert Opin. Ther. Pat. 21 (11) (2011) 1657–1672.

[11] K. Jagadish, J.A. Camarero, Cyclotides, a promising molecular scaffold for peptide-based therapeutics, Biopolymers 94 (5) (2010) 611–616.

[12] M. Maqueda, A. Galvez, M.M. Bueno, M.J. Sanchez-Barrena, C. Gonzalez, A. Albert, M. Rico, E. Valdivia, Peptide AS-48: prototype of a new class of cyclic bacteriocins, Curr. Protein Pept. Sci. 5 (5) (2004) 399–416.

[13] C.P. Scott, E. Abel-Santos, M. Wall, D.C. Wahnon, S.J. Benkovic, Production of cyclic peptides and proteins in vivo, Proc. Natl. Acad. Sci. U. S. A. 96 (24) (1999) 13638–13643.

[14] J.S. Zheng, S. Tang, Y. Guo, H.N. Chang, L. Liu, Synthesis of cyclic peptides and cyclic proteins via ligation of peptide hydrazides, Chembiochem 13 (4) (2012) 542–546.

[15] D.J. Craik, R.J. Clark, N.L. Daly, Potential therapeutic applications of the cyclotides and related cystine knot mini-proteins, Expert Opin. Investig. Drugs 16 (5) (2007) 595–604.

[16] D.J. Craik, M. Cemazar, N.L. Daly, The cyclotides and related macrocyclic peptides as scaffolds in drug design, Curr. Opin. Drug Discov. Devel. 9 (2) (2006) 251–260.

[17] C.W. Gruber, A.G. Elliott, D.C. Ireland, P.G. Delprete, S. Dessein, U. Göransson, M. Trabi, C.K. Wang, A.B. Kinghorn, E. Robbrecht, D.J. Craik, Distribution and evolution of circular miniproteins in flowering plants, Plant Cell 20 (9) (2008) 2471–2483.

[18] Q. Kaas, D.J. Craik, Analysis and classification of circular proteins in CyBase, Biopolymers 94 (5) (2010) 584–591.

[19] J.J. Ward, J.S. Sodhi, L.J. McGuffin, B.F. Buxton, D.T. Jones, Prediction and functional analysis of native disorder in proteins from the three kingdoms of life, J. Mol. Biol. 337 (2004) 635–645.

[20] B. Xue, A.K. Dunker, V.N. Uversky, Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life, J. Biomol. Struct. Dyn. 30 (2012) 137–149.

[21] M. Piippo, N. Lietzén, O.S. Nevalainen, J. Salmi, T.A. Nyman, Pripper: prediction of caspase cleavage sites from whole proteomes, BMC Bioinformatics 11 (2010) 320.

[22] I. Bertini, L. Decaria, A. Rosato, The annotation of full zinc proteomes, J. Biol. Inorg. Chem. 15 (7) (2010) 1071–1078.

[23] W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics 22 (2006) 1658–1659.

[24] H.M. Berman, et al., The protein data bank, Nucleic Acids Res. 28 (2000) 235–242.

[25] H. Zhang, T. Zhang, K. Chen, S. Shen, J. Ruan, L. Kurgan, Sequence based residue depth prediction using evolutionary information and predicted secondary structure, BMC Bioinformatics 9 (2008) 388.

[26] M. Mizianty, W. Stach, K. Chen, K.D. Kedarisetti, F. Miri Disfani, L. Kurgan, Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources, Bioinformatics 26 (18) (2010) i489–i496.

[27] P. Baldi, S. Brunak, Y. Chauvin, C. Andersen, H. Nielsen, Assessing the accuracy of prediction algorithms for classification: an overview, Bioinformatics 16 (2000) 412–424.

[28] K. Tomii, M. Kanehisa, Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins, Protein Eng. 9 (1996) 27–36.

[29] D.W. Buchan, S.M. Ward, A.E. Lobley, T.C. Nugent, K. Bryson, D.T. Jones, Protein annotation and modelling servers at University College London, Nucleic Acids Res. 38 (2010) W563–W568.

[30] H. Zhang, T. Zhang, K. Chen, K.D. Kedarisetti, M.J. Mizianty, Q. Bao, W. Stach, L. Kurgan, Critical assessment of high-throughput standalone methods for secondary structure prediction, Brief. Bioinform. 12 (6) (2011) 672–688.

[31] M.J. Mizianty, L. Kurgan, Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences, BMC Bioinformatics 10 (2009) 414.

[32] L. Kurgan, K. Cios, K. Chen, SCPRED: accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences, BMC Bioinformatics 9 (2008) 226.

[33] D.J. Craik, N.L. Daly, J. Mulvenna, M.R. Plan, M. Trabi, Discovery, structure and biological activities of the cyclotides, Curr. Protein Pept. Sci. 5 (2004) 297–315.

[34] X. Wu, et al., Top 10 algorithms in data mining, Knowl. Inform. Syst. 14 (1) (2008) 1–37.

[35] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.

[36] T.F. Smith, M.S. Waterman, Identification of common molecular subsequences, J. Mol. Biol. 147 (1981) 195–197.

[37] O. Gotoh, An improved algorithm for matching biological sequences, J. Mol. Biol. 162 (1982) 705–708.

[38] UniProt Consortium, Reorganizing the protein space at the Universal Protein Resource (UniProt), Nucleic Acids Res. 40 (2012) D71–D75.

[39] L.Y. Geer, A. Marchler-Bauer, R.C. Geer, L. Han, J. He, S. He, C. Liu, W. Shi, S.H. Bryant, The NCBI BioSystems database, Nucleic Acids Res. 38 (2010) D492–D496.

[40] F.D. Ciccarelli, T. Doerks, C. von Mering, C.J. Creevey, B. Snel, P. Bork, Toward automatic reconstruction of a highly resolved tree of life, Science 311 (2006) 1283–1287.