Journal of
CHEMICAL INFORMATION
AND MODELING

ARTICLE

pubs.acs.org/jcim

CycloPs: Generating Virtual Libraries of Cyclized and Constrained Peptides Including Nonnatural Amino Acids

Fergal J. Duffy,*[†] Mélanie Verniere,[†] Marc Devocelle,[‡] Elise Bernard,[§] Denis C. Shields,[†] and Anthony J. Chubb[†]

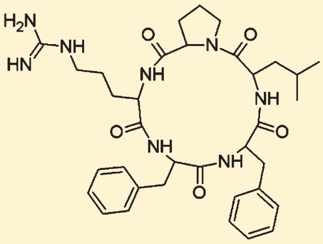
[†]UCD Complex and Applied Systems Laboratory, Conway Institute of Biomolecular and Biomedical Sciences, and School of Medicine and Medical Sciences, University College Dublin, Belfield, Dublin 4, Ireland

[‡]Department of Pharmaceutical & Medicinal Chemistry, Centre for Synthesis and Chemical Biology, Royal College of Surgeons Ireland, 123 St. Stephens Green, Dublin 2, Ireland

[§]Laboratory of Molecular Biology, Medical Research Council, Hills Road, Cambridge CB2 0QH, United Kingdom

S *Supporting Information*

ABSTRACT: We introduce CycloPs, software for the generation of virtual libraries of constrained peptides including natural and nonnatural commercially available amino acids. The software is written in the cross-platform Python programming language, and features include generating virtual libraries in one-dimensional SMILES and three-dimensional SDF formats, suitable for virtual screening. The stand-alone software is capable of filtering the virtual libraries using empirical measurements, including peptide synthesizability by standard peptide synthesis techniques, stability, and the druglike properties of the peptide. The software and accompanying Web interface is designed to enable the rapid generation of large, structurally diverse, synthesizable virtual libraries of constrained peptides quickly and conveniently, for use in virtual screening experiments. The stand-alone software, and the Web interface for evaluating these empirical properties of a single peptide, are available at http://bioware.ucd.ie.



INTRODUCTION

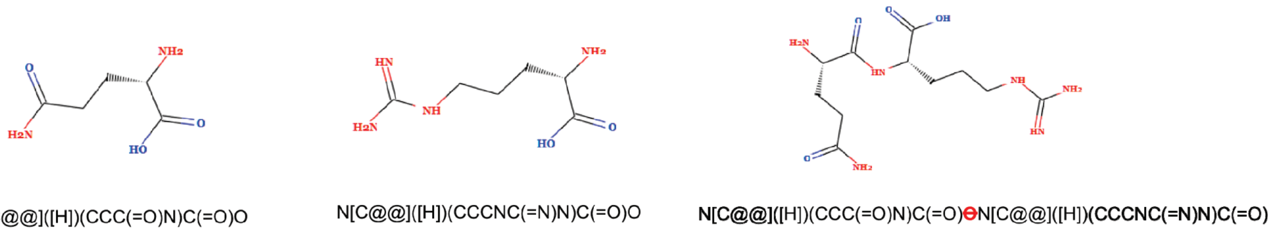
Constrained peptides comprise a wide variety of conformationally limited peptides. Conformational constraints include disulfide bonds between two cysteine residues, head—tail bonded peptides, and other forms of peptide side-chain to backbone and peptide side-chain to side-chain bonds. Constrained peptides may act as peptidomimetics of natural, linear peptides,[1,2] and may make more suitable therapeutic candidates, as the conformational constraint has been seen to increase specificity to a protein target,[3] as well as provide protection from proteolytic degradation.[4]The advantages that constrained peptides display over linear peptides have led to their use as potential protein binding compounds in phage-display libraries,[5] as mimics of protein turn structure in virtual libraries,[6] and as binding ligands of RNA in virtual libraries.[7]

To generate virtual libraries of constrained peptides, it is necessary to use chemoinformatics techniques to assemble the amino acid building blocks into a linear peptide. The SMILES[8] (Simplified Molecular Input Line Entry System) chemical notation system may be used to computationally assemble constrained peptides as a string of text, unambiguously describing each atom and bond in the molecule in a manner amenable to machine processing. The SMILES string can be used as a basis to generate 2D and 3D depictions of molecules,[9] and several open-source and commercial packages[10−14] exist with this capability. Three-dimensional coordinates are usually generated using molecular mechanics methods and are commonly written as data to chemical table files[9] such as the MDL Molfile, or MDL structure-data file (.sdf file), which hold coordinates and connection tables for one or more molecules. The open-source RDKit[10] toolkit provides resources to generate 3D and 2D coordinates for virtual library molecules.

In addition to the natural amino acids, there is a large range of unusual amino acids with various interesting properties which are commercially available and directly utilizable in peptide synthesis by the Fmoc—*t*-Bu strategy.[15] It makes sense to allow for their inclusion in any virtual peptide library. The ZINC[16] database is a noncommercial database which, as of this writing, contains over 20 million purchasable compounds in a variety of formats suitable for virtual screening, including a number of nonnatural amino acids.

While a number of commercially available software packages provide programs that may be coded together to generate virtual peptide libraries, some of these require organizational management of the steps that yield a pipeline of library synthesis, and some are quite expensive for academic users. CycloPs offers a graphical method of assembling and refining libraries of constrained peptides. The motivation behind CycloPs is to provide a fast and convenient way to create virtual libraries of constrained peptides for virtual screening, with wide structural diversity, all readily synthesizable using standard peptide synthesis

Received: October 28, 2010
Published: March 24, 2011



N[C@@]([H])(CCC(=O)N)C(=O)O

N[C@@]([H])(CCCNC(=N)N)C(=O)O

N[C@@]([H])(CCC(=O)N)C(=O)⊖N[C@@]([H])(CCCNC(=N)N)C(=O)

Figure 1. Formation of the peptide bond between glutamic acid and asparagine, resulting in the dipeptide plus water that can be mimicked by appropriately arranged SMILES strings. The SMILES strings are concatenated, while removing the "O" representing the C-terminal hydroxyl group in the SMILES string.

Table 1. Constraint Types Recognized by CycloPs

name	residues involved	description	CycloPs abbreviation
head-to-tail bond	C-terminal and N-terminal amino acids	peptide bond between N-terminal amino group and C-terminal carboxyl group	HT
disulfide bond	pair of cysteine residues	sulfur−sulfur bond between side chains of cysteine residues	SS
side-chain to side-chain bond	lysine to glutamic acid or aspartic acid serine or threonine to aspartic acid or glutamic acid (depsipeptide)	amide (−CONHR−) or depsipeptide (−COOR−) bond between side chains of residues with an amine or hydroxyl group and residues with a carboxyl group	SC or SCSC
side-chain to N-terminal bond	glutamic or aspartic acid to peptide N-terminal	amide bond between residue carboxyl side-chain group and N-terminal amine group	SC or SCNT
side-chain to C-terminal bond	lysine, serine, or threonine to peptide C-terminal	amide or depsipeptide bond between residue amine or hydroxyl side-chain group and C-terminal carboxyl group	SC or SCCT

methods.[15] The final goal is to identify potentially therapeutic, protease resistant, small cyclized peptides that can be synthesized and tested relatively cheaply and quickly. This is accomplished by including nonnatural amino acids, and allowing the user to filter by the druglike (compliance with Lipinski's "rule-of-five") properties of the constrained peptide and the synthesizability/stability of the peptide. CycloPs is designed to support the generation of virtual libraries of arbitrary size, limited only by the disk space of the user, and therefore is written in a memory-efficient way in the Python programming language.

■ IMPLEMENTATION AND RESULTS

The functionality of CycloPs is broken into two main parts: (i) an interface for generating and examining a single constrained peptide, and (ii) an interface for generating virtual libraries of constrained peptides. The software is also provided as a Web application which mimics the functionality of the single peptide interface at http://bioware.ucd.ie/. The ZINC nonnatural amino acids are not included in the Web interface.

Generating Single Constrained Peptides. *Assembling the Linear Peptide.* At the heart of CycloPs is the ability to assemble constrained peptides from amino acids using SMILES structures. A virtual peptide can be assembled from virtual amino acids in SMILES form in a manner analogous to the formation of the peptide bond in a physical peptide. See Figure 1.

The formation of the peptide SMILES string from the individual amino acid strings exploited the property of the SMILES encoding that there are many ways to write the same molecule, and if the SMILES string is arranged to start with the N-terminus and end with the C-terminus, the amount of manipulation required to combine the two amino acids into a peptide is minimized. CycloPs stores an internal library of common amino acids in SMILES format to facilitate rapid peptide structure assembly.

Constraining the Peptide. The stored SMILES strings can very easily be used to generate a single peptide, but generating a constrained peptide adds an additional complication. CycloPs supports several varieties of structural constraint, which generally result in a linear peptide being bonded back onto itself to form a large central macrocycle. Table 1 lists descriptions of each type of constraint supported by CycloPs and the abbreviations that CycloPs uses internally and writes to output files to label peptides constrained in a particular way. Figure 2 illustrates these constraints.

The suitability of each individual amino acid for a particular cyclization and the atoms involved in the resulting chemical bond are predefined in CycloPs's internal library of common natural amino acid SMILES strings. While working with a single peptide, CycloPs takes as input a linear peptide sequence, analyzes it for applicable constraint methods, and allows the user to select which constraint, if any, is desired.

CycloPs places a few limitations on which constraints may be chosen, to protect against generation of virtual peptides that may be sterically hindered or too highly structurally strained to be feasibly synthesized. These empirical rules are probably both incomplete and excessively rigorous, but they are included as an optional filter to facilitate the efficient synthesis of candidate peptides. Head-to-tail bonded peptides must have at least five residues, except the two residue diketopiperazines. A spacing of at least two residues is necessary between side chain and side chain, or side chain and peptide termini. The complete list of restrictions is outlined in Table 2.

Output. On specification of the desired peptide sequence and constraint, CycloPs will generate a virtual peptide in SMILES format, which explicitly specifies the structure of the molecule. This can then be translated to a 2D or 3D representation of the peptide using the RDKit toolkit.[10]

Using the SMILES string, it is possible to evaluate properties of the molecule by analyzing its structure. CycloPs displays, along with a 2D representation of the input peptide, the log *P* value of the peptide, the number of hydrogen bond donors and acceptors, and the molecular weight of the peptide. All of these properties are calculated using the RDKit.

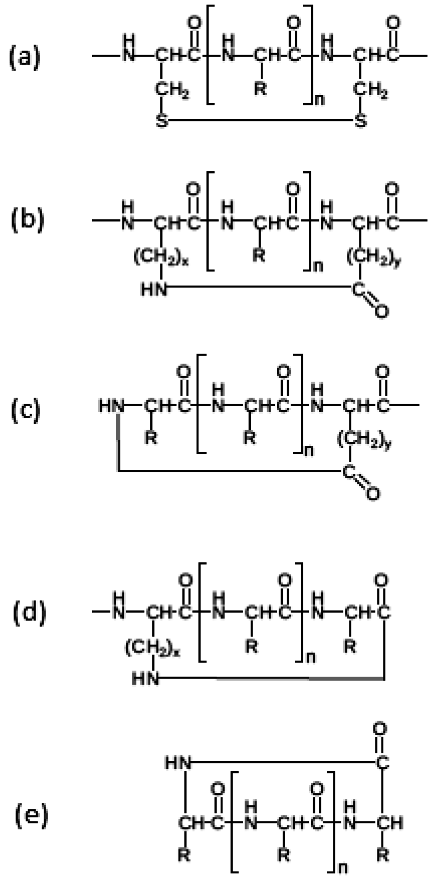


Figure 2. Chemical structure of each type of peptide structural constraint supported by CycloPs. (a) Disulfide bonded peptide with two cysteine residues. (b) Side-chain to side-chain bonded peptide with side chains containing amino, carboxyl, or hydroxyl groups, which cyclize the peptide forming amide or ester bonds. (c) Side-chain to N-terminal bond. (d) Side-chain to C-terminal bond. (e) Head-to-tail bond.

CycloPs also displays the results of some empirical tests of the linear peptide, related to its synthesizability by routine peptide synthesis techniques.[17] These include properties such as the logarithm of the octanol/water partition coefficient (log *P*(o/w)), calculable from the peptide SMILES string.[19,20] These rules are summarized in Table 3.

For the user, the single peptide to be generated can be specified using the single letter amino acid code, for the natural amino acids. CycloPs has restricted the use of upper and lower cases, using only upper case letters for L-amino acids and only lower case letters for D-amino acids (such as PQPL or dAiW). Amino acids may also be specified using the full name or ZINC reference number of the peptide, separated by commas (e.g., L-arginine, D-cysteine, glycine, ZINC00388696). Figure 4 displays the actual interface of the single peptide part of the program, as seen on a Microsoft Windows machine. A mixture of single letter and peptide names is not allowed (e.g., AAdT, ZINC00388696, TS is not a valid sequence).

Web Server. Most of the single peptide functionality of CycloPs is also available, for convenience, as a Web server. The virtual library generation features of CycloPs are not available on the Web server, because generating a large library is computationally expensive, as well as generating very large files, and is best done locally.

Generating Virtual Libraries of Constrained Peptides. *Library Definition.* CycloPs is designed to allow the generation of arbitrarily large libraries of peptide structures.

CycloPs generates virtual libraries combinatorially, iterating through all possible peptides of a specified number of residues. To design a library, the user can specify exactly which amino acids should be included, which constraints (including linear peptides as an option) should be included, and whether to include or exclude peptides based on hydrophobicity, and druglike properties.

To create a peptide library, a library pattern must be specified. The pattern is a string of text, with the standard letter characters representing amino acids, and "X" representing a variable position. Every peptide that matches a particular pattern is generated as part of the library.

Thus, KXXDX would represent a virtual library pattern where the first and fourth positions are fixed as L-lysine and L-aspartic acid respectively, and every other position could be any defined amino acid. When generating the library, KXXDX will be

Table 2. Limitations on Peptides Suitable for Cyclization

constraint	limitations	examples of valid sequences
head-to-tail bond	sequence must be either five or more amino acids in length, or exactly two amino acids (not three or four)	MNSYT, KA, HRKCNVGD
side-chain to side-chain bond	sequence must have a spacing of at least two[a] amino acids between bonded side chains	KAAD, ESCCE
side-chain to N-terminal bond	sequence must have a spacing of at least two[a] amino acids between bonded side chain and N-terminus	RFACGE, ASD
side-chain to C-terminal bond	sequence must have a spacing of at least two[a] amino acids between bonded side chain and C-terminus	KAS, RSEGE
disulfide bond	sequence must have a spacing of at least two[a] amino acids between bonded cysteine side chain residues	CAAC, FACERC

[a] The rule of two amino acids minimum between the residues involved in the cyclization for this constraint is implemented on the basis of convenient and rapid synthesis of the peptides, without requiring repeated attempts; this does not constitute an absolute rule of synthetic feasibility.[1]

expanded into every possible matching peptide. By default, CycloPs begins with a working library of the 20 natural amino acids, but it can be expanded to include any of the enantiomeric D-amino acids (or reduced to exclude any natural amino acids), along with amino acids defined in the ZINC database.

With standard settings, the pattern KXXDX will expand to 20 × 20 × 20 = 8000 linear peptides.

Next, each linear peptide is tested against the panel of constraint types specified for the particular virtual library. If a peptide is compatible with one or more constraints, the resulting constrained peptides are added to the library. Table 4 shows a selection of linear peptides and the constrained peptides that may be generated from them.

Table 3. Properties of a Difficult Peptide

category	details
forbidden[a] amino acid motifs	no more than two consecutive prolines
	Asp-Gly and Asp-Pro are forbidden
	Asp or Gln forbidden at (unmodified) N-terminus
physical properties	log P(o/w) value >0 (peptide is hydrophobic)
	less than one charged residue (lysine, arginine, histidine, aspartic acid, glutamic acid) every five residues

[a] Motifs that have a potential synthetic or stability concern related to the sequence.

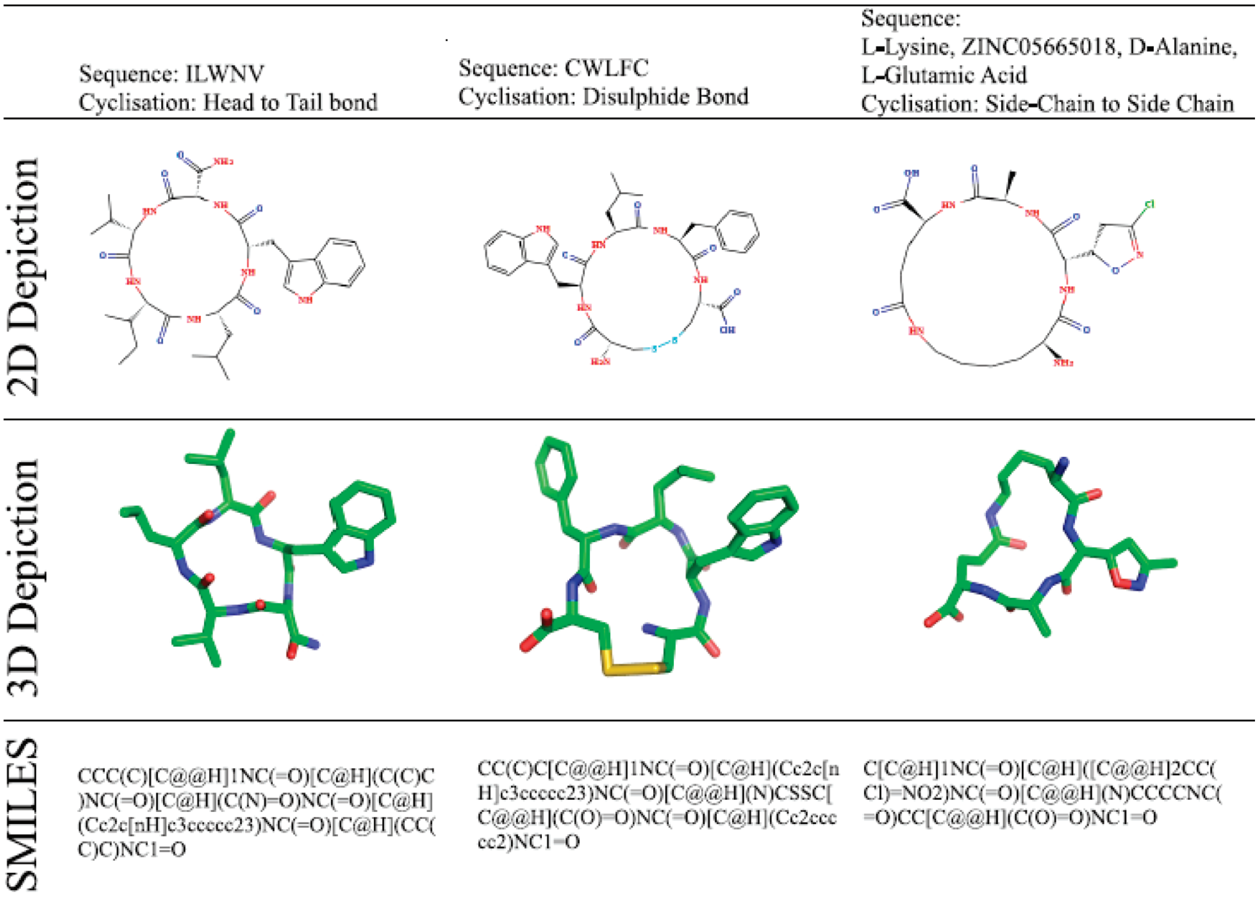
Library Filtering and Restriction. When designing combinatorial virtual libraries, the combinatorial explosion of possible structures is a serious and much discussed issue, and putting together the optimal library from a diverse set of building blocks is a difficult problem.[18] CycloPs incorporates methods both to reduce the set of building blocks and to filter out undesirable structures from the library as it is generated.

The working library is, by default, made up of all 20 natural amino acids. Also available are the D-amino acids, a few selected common amino acids (such as L-ornithine), and commercially available amino acid structures from ZINC. Any of these amino acids can be added to or removed from the library.

To reduce the number of unpromising peptides from the library, CycloPs allows difficult-to-synthesize peptides to be excluded from the library during generation. The rules determining a "difficult" peptide have been summarized in Table 3.

CycloPs can also exclude peptides that are not druglike according to Lipinski's "rule-of-five".[21] This may be of somewhat limited use for peptides longer than two to three residues, as they will almost certainly fail these criteria.

Output. When outputting the virtual library, CycloPs will attempt to name output files in a logical manner. Compounds are generally named for their sequence and constraint: if they are made up of the basic 20 natural L-amino acids, or their D-amino equivalents, letters will be used for the sequence; otherwise each amino acid name, separated by a comma, will be used. The constraint will be appended to the sequence as its CycloPs



Sequence: ILWNV
Cyclisation: Head to Tail bond

Sequence: CWLFC
Cyclisation: Disulphide Bond

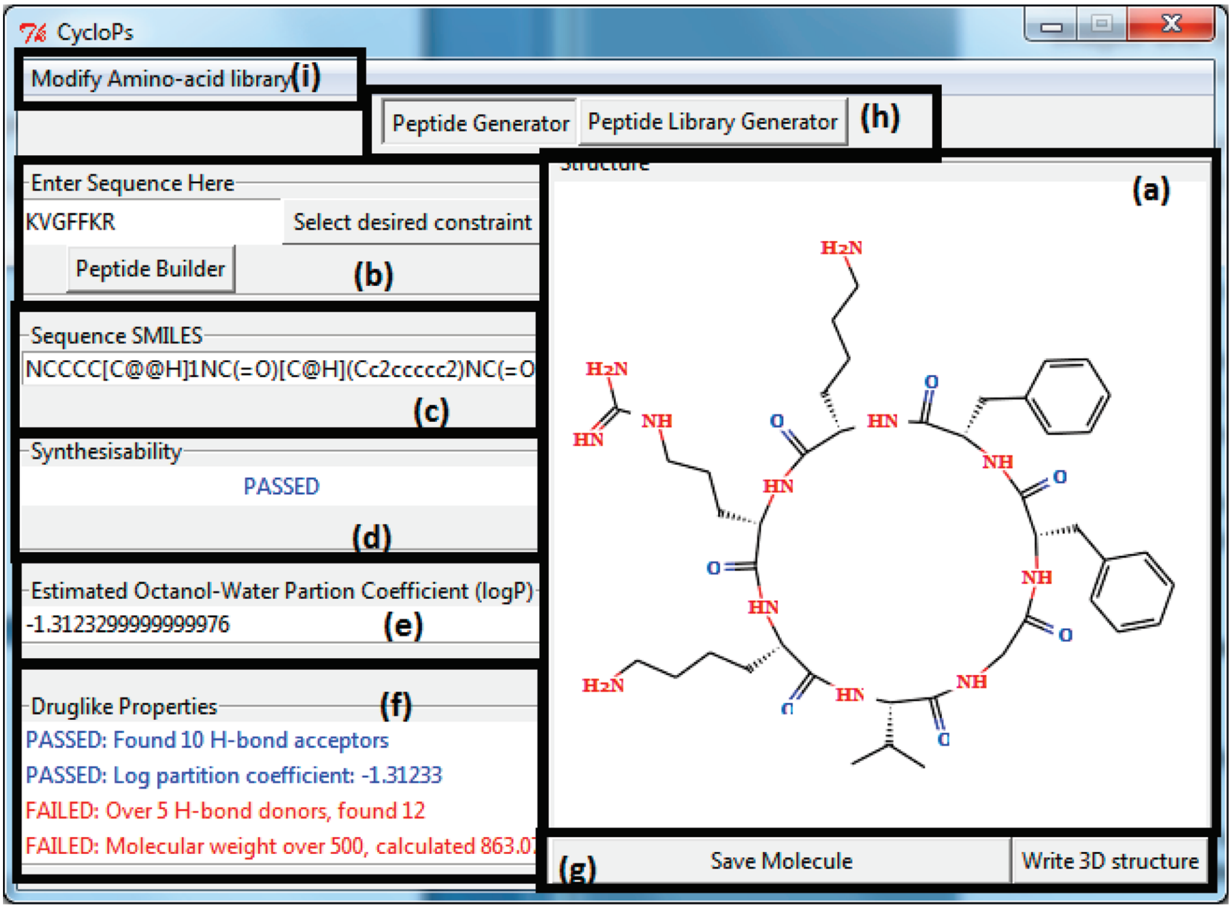
Sequence:
L-Lysine, ZINC05665018, D-Alanine, L-Glutamic Acid
Cyclisation: Side-Chain to Side Chain

CCC(C)[C@@H]1NC(=O)[C@H](C(C)C)NC(=O)[C@H](C(N)=O)NC(=O)[C@H](Cc2c[nH]c3ccccc23)NC(=O)[C@H](CC(C)C)NC1=O

CC(C)C[C@@H]1NC(=O)[C@H](Cc2c[nH]c3ccccc23)NC(=O)[C@@H](N)CSSC[C@@H](C(O)=O)NC(=O)[C@H](Cc2ccccc2)NC1=O

C[C@H]1NC(=O)[C@H]([C@@H]2CC(Cl)=NO2)NC(=O)[C@@H](N)CCCCNC(=O)CC[C@@H](C(O)=O)NC1=O

Figure 3. Examples of peptide depiction output by CycloPs. The three peptides included here demonstrate three different types of constrained peptides, in 1D (SMILES), 2D, and 3D forms for each peptide. The peptide on the right also includes a nonnatural amino acid taken from the ZINC database.



KVGFFKR

NCCCC[C@@H]1NC(=O)[C@H](Cc2ccccc2)NC(=O

PASSED

-1.3123299999999976

PASSED: Found 10 H-bond acceptors

PASSED: Log partition coefficient: -1.31233

FAILED: Over 5 H-bond donors, found 12

FAILED: Molecular weight over 500, calculated 863.07

Figure 4. CycloPs single peptide interface. This figure shows how CycloPs' single peptide interface is arranged using lettered boxes. (a) indicates the canvas displaying the currently generated peptide. (b) shows the peptide sequence entry box and the constraint selection menu. The Peptide Builder button allows the entry of long sequences of ZINC amino acids without typing in each full ZINC reference. (c) is where CycloPs displays the SMILES string of the current peptide. (d) indicates whether the peptide passes the synthesizability rules mentioned above, and detailed in Table 4. (e) gives the estimated log P(o/w) value: a measure of the molecular hydrophobicity. (f) details the druglike (Lipinski's "rule-of-five") properties of the peptide: blue for druglike; red for not druglike. (g) is where the 2D or 3D depiction can be saved to a file. (h) allows a choice between the single peptide and library generation panes, and (i) is the menu where the included amino acids may be selected.

abbreviation, and disulfide and side-chain bonded peptides will have the bonding residues indicated.

For example, KASD-SCNXXZ indicates a peptide of sequence KASD, with a side-chain to side-chain bond of an N-terminal-like pattern (K) to a C-terminal-like residue (D). CQPC-SSCXXC indicates a peptide of sequence CQPC with a disulfide bond between the two cysteines.

In the constraint pattern, "X" represents a residue not involved in the cyclization, "Z" represents a C-terminal-like residue side chain involved in cyclization, "E" represents a residue side chain with a hydroxyl group involved in an ester depsipeptide bond, and "N" represents an N-terminal-like side chain involved in cyclization. "C" represents a cysteine residue in a disulfide bond, but it could also be used for a nonnatural amino acid if that was suitable for disulfide binding. Figure 3 gives sample depictions of a number of constrained peptides generated by CycloPs.

If the virtual library is output as a single .sdf file of many 3D peptide structures, the names will be included as molecule properties within the file. If the library is output as 2D .png images, each image will be named after its molecule. If the library is output in SMILES format, it will be written as a comma separated value (.csv) file with names in the first column and SMILES in the second. Figure 5 displays an annotated version of the virtual library generation control window of the program, similar to Figure 4.

Nonnatural Amino Acids. To further extend the range of possible amino acids that could be incorporated into the CycloPs peptide generator, we developed an automated system for downloading the structures of suitable purchasable compounds from the ZINC Web site[16] and converting them into a compatible format. This is illustrated in Scheme S1 in the Supporting Information.

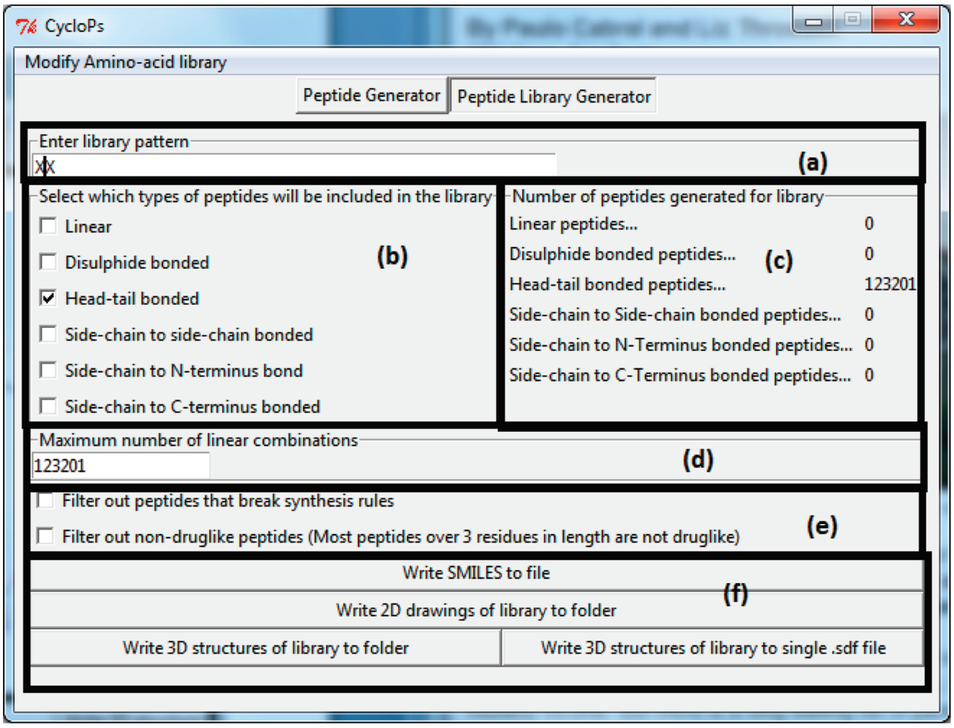
Filtering for Fmoc Containing Amino Acids. The ZINC subset no. 6 "all-purchasable", updated June 17, 2010, contains the molecular structures of 20 669 372 molecules that have available vendor details (see Table 5). The minimum requirement for peptide synthesis according to the Fmoc—*t*-Bu strategy (the most widely used method for peptide synthesis[22]) is an N-terminal fluorenylmethyloxycarbonyl (Fmoc) group, and a free carboxylate group.[18] After specifying the location (URL) of the requisite data set, we downloaded the data set and used RDKit functions to perform a 2D search of each molecule to find the minimal "Fmoc—N" and "COOH" features needed to identify potential amino acids. This was further broken down into a two-stage process, as the "COO" motif of the carboxylate group is contained in the carbamate motif of the Fmoc. The first step was thus to search for "Fmoc—N—C", which returned 1217 molecules, followed by replacement of the Fmoc moiety with an asterisk handle in the SMILES string. These compounds were then searched for a free carboxylate, leaving 849 potential amino acids. To avoid a possible complication with preactivated carboxylate groups, we searched the database for activated carboxylates. None were found.

Removing Protecting Groups. Peptide synthesis involves the cyclical removal of the protecting Fmoc moiety from the growing peptide chain, followed by coupling of the next Fmoc-containing amino acid.[18] The side chains of amino acids that contain oxygen, nitrogen, or sulfur thus need to be blocked during these cycles, or else nonspecific addition in branched chains would result. These blocking groups are generally removed by treatment with TFA (trifluoroacetic acid) during the isolation step of the assembled sequence. To model this chemistry, the side chain protecting groups need to be removed in the same manner as would occur during TFA cleavage. Protecting groups were removed in descending size, to avoid the possibility of a protecting group substructure being removed before the complete protecting group (see Chart S1 in the Supporting Information for a list of all protecting groups). Duplicate molecules were also removed.

Rearranging SMILES. The input format for the CycloPs peptide generator is a list of amino acids in SMILES format starting from the Fmoc-amino end and finishing with the free carboxylate. As SMILES are a one-dimensional representation of a three-dimensional graph, it is independent of the starting point and order of substituents. Thus some of the ZINC SMILES were in the correct orientation (302), but many were not (547). We developed a randomization strategy to reorder the SMILES in the correct orientation. Critically, this involved both fractionating the SMILES string into the correct substituents and using a comparison function in RDkit to compare the new randomized structure to the original molecule. As the comparator function compares the three-dimensional graph format of the new molecule to the original molecule, it is independent of the orientation of the input SMILES sequence.

Table 4. Examples of Suitable Constraints for Various Linear Peptides

linear peptide	valid constraints
KAAD	side-chain to side-chain bond (L-lysine to L-aspartic acid) side-chain to C-terminal bond (L-lysine to C-terminus) side-chain to N-terminal bond (L-aspartic acid to N-terminus)
CKLMC	disulfide bond (L-cysteine to L-cysteine) head-to-tail bond (N-terminus to C-terminus, and is over four residues in length)
L-threonine, D-cysteine, ZINC05665018, L-arginine, L-cysteine	disulfide bond (D-cysteine to L-cysteine) head-to-tail bond (N-terminus to C-terminus, and is over four residues in length) side-chain to C-terminal bond (depsipeptide L-threonine to C-terminus)


CycloPs

Modify Amino-acid library

Peptide Generator | Peptide Library Generator

Enter library pattern: XX

Select which types of peptides will be included in the library
- ☐ Linear
- ☐ Disulphide bonded
- ☑ Head-tail bonded
- ☐ Side-chain to side-chain bonded
- ☐ Side-chain to N-terminus bond
- ☐ Side-chain to C-terminus bonded

Number of peptides generated for library
- Linear peptides... 0
- Disulphide bonded peptides... 0
- Head-tail bonded peptides... 123201
- Side-chain to Side-chain bonded peptides... 0
- Side-chain to N-Terminus bonded peptides... 0
- Side-chain to C-Terminus bonded peptides... 0

Maximum number of linear combinations: 123201

- ☐ Filter out peptides that break synthesis rules
- ☐ Filter out non-druglike peptides (Most peptides over 3 residues in length are not druglike)

(a)
(b)
(c)
(d)
(e)
(f)
Write SMILES to file
Write 2D drawings of library to folder
Write 3D structures of library to folder
Write 3D structures of library to single .sdf file

Figure 5. CycloPs library generation interface. (a) Library pattern entry box. (b) Library included constraint selection pane. (c) Library peptide count information pane. (d) Maximum number of linear combinations: this gives a quick estimate of the library size. If this is above 1 000 000 peptides, the counts for each constraint will not be generated due to time constraints, as each sequence is generated and tested individually for each constraint. (e) Library filtering options. (f) Library output options.

Table 5. Filtering the ZINC Database for Amino Acids

subset	number of molecules
ZINC subset no. 6	20 669 372
Fmoc−N−C only	1217
one free COOH	849
unique nonnatural amino acids	347

For example, the ZINC molecule ZINC_4521508 has the following starting SMILES: “CCCC[C@H](C(═O)[O−])N[*]”, where [*] replaces the N-terminal Fmoc moiety. To reorder this SMILES in the format N−X−C, where “N” is the N-terminus, “C” is the C-terminus, and “X” is everything in between, we split the SMILES into “[*]N”, “C(═O)[O−]”, and “CCCC[C@H]( )”. The variable region is then further subdivided into a list of component parts, (“C”, “C”, “C”, “C”, “[C@H]”, “(”, “)”), which are recombined randomly. The number of parentheses pairs are also randomized, as these define branch points which are altered by the reorientation of the SMILES notation. To preserve chirality, we also randomized the number of “@” symbols. Thus the “[C@H]” in this example could also be written “[C@@H]” in the randomly generated SMILES. At each randomization cycle the output SMILES structure was compared to the original input SMILES structure, allowing chiral-specific comparison between both SMILES. This relies on the fact that although there exist many possible SMILES strings to represent one molecule, these are all one-dimensional representations of the same three-dimentional interconnection between various atoms. We exploited an algorithm in RDKit that calculates a single, canonical SMILES[23] for any one molecule. The program can thus directly compare the unique SMILES generated for both the input molecule and the hypothetical molecule generated by the random rearrangement. The programming loop ended when a match was found, or else a cutoff of 15 000 cycles was used to discard the molecule. For instance, in the above example the output SMILES was “[*]N[C@H](CCCC)C(═O)[O−]”. Ninety molecules were reordered this way, yielding a total of 392 molecules in the correct SMILES orientation. These were then screened for duplicates, resulting in 347 nonnatural amino acids that can be incorporated into CycloPs.

Integration into CycloPs. In addition to the 20 natural amino acids, this survey of ZINC provided an additional 347 nonnatural amino acids that can be used in peptide synthesis. These were incorporated into the CycloPs peptide generator by saving the text file in the relevant directory, which is read at run time by CycloPs. This file can be periodically updated at the Cyclops download site and server, and in addition the file may be edited by users of the Cyclops software (file name zinc_output). The names and SMILES strings for these nonnatural amino acids are included in Table S1 in the Supporting Information.

DISCUSSION AND CONCLUSIONS

Peptides may be naturally attractive mimetics of protein interactions. However, computational searches with constrained peptides are likely to more accurately reflect the in vitro pose than that of less constrained linear peptides. The reduced conformational space of cyclized constrained peptides is less computationally demanding and more easily replicated in chemical synthesis. Furthermore, affinity will be improved by constraining the peptide in the active conformation. Thus, we were motivated to streamline the design of virtual constrained peptide libraries.

Generating libraries of molecules for virtual screening is a common chemoinformatics workflow. At the moment, it is possible to generate these libraries, either paying for a commercial package or spending time developing and debugging scripts to tie a variety of open-source tools into the desired workflow. The creation of CycloPs was motivated by a desire to make generating these libraries of constrained peptides easy and flexible, enabling more time to be spent analyzing results and less time setting up experiments. CycloPs is freely available for any use by those interested in constrained peptides. It is written in a memory-efficient way, allowing the generation of arbitrarily large virtual libraries, and it is limited only by disk space. It has been tested for generating SMILES libraries of over 10 million peptide structures.

The ability to include nonnatural amino acids in CycloPs generated libraries adds a huge amount of potential structural diversity to constrained peptide libraries, especially considering that the nonnatural amino acid structures available from ZINC vastly outnumber the natural amino acids (over 300 versus 20). The increased range of structural diversity in the ZINC amino acids can enable improvements in virtual screening efficiency. One method of improving the speed and performance of computational docking of cyclized peptides is to limit the number of rotational bonds in the constituent amino acids; however, this will disproportionately remove positively charged amino acids, as both lysine and arginine have long, flexible side chains. Within the diversity of nonnatural amino acids, it is likely that a suitable positively charged amino acid with few rotational bonds in the side chain can be added to the library to compensate, for example, ZINC19363627 in place of L-lysine. In this way, constrained peptide virtual libraries can be tuned for the form of screening which will later be used to match them to a target.

Chemoinformatics screening methods are quite well developed, and they include ligand based[24−26] and 3D structure based pharmacophore screening,[25] and computationally docking[27] a ligand into the (NMR or X-ray crystallographically determined) ligand binding 3D structure of the protein target of interest. The variety of output formats from CycloPs lends itself to this analysis, with SMILES format being quickly and easily parsed into a 2D pharmacophore fingerprint for comparison with the corresponding pharmacophore fingerprints of one or more known true ligands. Three-dimensional “.sdf” structures can be compared to a 3D pharmacophore generated from the structure of the protein target binding pocket, or docked into the binding pocket directly.

After computational screening, synthesis of selected compounds is relatively straightforward. The methods for chemically synthesizing constrained peptides are also well understood and developed. They can be synthesized as linear peptides by solid-phase synthesis, before selectively removing protecting groups on side chains or terminals involved in binding to allow formation of the final cyclical constrained peptide.

Virtual libraries of constrained peptides are also of interest in phage display analyses. The likely structural conformation of the peptides generated in phage display may be generated using CycloPs. Computationally it may be unfeasible to dock or pharmacophore match all the virtual library for a large phage display, but it would be possible to dock both the positive hits from phage display and a randomly selected subset of the virtual library as a control set, to determine the computational matching of the phage display findings to the target.

Another potentially useful application of CycloPs is the semiautomated design of novel inhibitors of protein−protein interactions. Although protein−protein interactions are critical in cellular signaling cascades, and thus make promising targets for therapeutic drug discovery, the development of inhibitors is hampered by their large solvent-exposed surface areas. The near-exponential growth of the PDB protein structure database, coupled with advances in mass spectrometry and the characterization of interactomes, will result in numerous novel protein—protein interfaces involved in cell signaling. These binding sites can then be converted into pharmacophores and searched against libraries of constrained cyclic peptides produced by CycloPs, and may result in novel therapeutics.

Implementation and Availability. CycloPs in written in Python and executables are available, free for any commercial or noncommercial use, from http://bioinfo-casl.ucd.ie/cyclops/Download.html. A limited, Web server version of CycloPs is also available for working with single constrained peptides on the Web site at http://bioinfo-casl.ucd.ie/cyclops. CycloPs is distributed as a self-contained binary, and is available for Windows, Mac OS X, and Linux (tested on Ubuntu Linux).

ASSOCIATED CONTENT

S Supporting Information. SMARTS patterns for each protecting group removed from ZINC nonnatural amino acids are listed in Chart S1. The names and SMILES structure of each nonnatural amino acid gathered from ZINC are tabulated in Table S1. Scheme S1 describes processing of ZINC structures to find nonnatural amino acids. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: fergal.duffy@ucd.ie.

ACKNOWLEDGMENT

The authors thank Science Foundation Ireland for funding this work, and Niall Haslam for help with making the Web server version of CycloPs publicly available.

REFERENCES

(1) Hruby, V. J. Designing peptide receptor agonists and antagonists. *Nat. Rev. Drug Discovery* **2002**, *1* (11), 847–858.

(2) Vagner, J.; Qu, H.; Hruby, V. J. Peptidomimetics, a synthetic tool of drug discovery. *Curr. Opin. Chem. Biol.* **2008**, *12* (3), 292–296.

(3) Nam, N. H.; Ye, G.; Sun, G.; Parang, K. Conformationally constrained peptide analogues of pTyr-Glu-Glu-Ile as inhibitors of the Src SH2 domain binding. *J. Med. Chem.* **2004**, *47* (12), 3131–3141.

(4) Hanessian, S.; Auzzas, L. The practice of ring constraint in peptidomimetics using bicyclic and polycyclic amino acids. *Acc. Chem. Res.* **2008**, *41* (10), 1241–1251.

(5) Ladner, R. C. Constrained peptides as binding entities. *Trends Biotechnol.* **1995**, *13* (10), 426–430.

(6) Arbor, S.; Marshall, G. R. A virtual library of constrained cyclic tetrapeptides that mimics all four side-chain orientations for over half the reverse turns in the protein data bank. *J. Comput.-Aided Mol. Des.* **2009**, *23* (2), 87–95.

(7) Burns, V. A.; Bobay, B. G.; Basso, A.; Cavanagh, J.; Melander, C. Targeting RNA with cysteine-constrained peptides. *Bioorg. Med. Chem. Lett.* **2008**, *18* (2), 565–567.

(8) Weininger, D. SMILES, a Chemical Language and Information-System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31–36.

(9) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of Several Chemical-Structure File Formats Used by Computer-Programs Developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32* (3), 244–255.

(10) Landrum, G. *RDKit*, Q2 2010 1; Palo Alto, CA 94301, USA.

(11) *Molecular Operating Environment*, 2010.10; Chemical Computing Group: Montreal, Quebec H3A 2R7, Canada.

(12) *Discovery Studio*, 2.5; Accelrys: San Diego, CA 92121, USA; 2009.

(13) *DayCart*, 4.9; Daylight Chemical Information Systems: Laguna Niguel, CA 92677, USA; 2010.

(14) *OEChem*, 1.7.4; Open Eyes Scientific Software: Santa Fe, NM 87508, USA; 2010.

(15) Bray, B. L. Large-scale manufacture of peptide therapeutics by chemical synthesis. *Nat. Rev. Drug Discovery* **2003**, *2* (7), 587–593.

(16) Irwin, J. J.; Shoichet, B. K. ZINC—a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45* (1), 177–182.

(17) Coin, I.; Beyermann, M.; Bienert, M. Solid-phase peptide synthesis: from standard procedures to the synthesis of difficult sequences. *Nat. Protoc.* **2007**, *2* (12), 3247–3256.

(18) Brennan, M. P.; Cox, D.; Chubb, A. J. Peptide Diversity in Drug Discovery. *Front. Drug Des. Discovery: Struct.-Based Drug Des. 21st Century* **2007**, *3*, 395–432.

(19) Convard, T.; Dubost, J. P.; Le Solleu, H.; Kummer, E. SmilogP: A Program for a Fast Evaluation of Theoretical Log P from the Smiles Code of a Molecule. *Quant. Struct.-Act. Relat.* **1994**, *13* (1), 34–37.

(20) Klopman, G.; Li, J. Y.; Wang, S. M.; Dimayuga, M. Computer Automated Log P Calculations Based on an Extended Group-Contribution Approach. *J. Chem. Inf. Comput. Sci.* **1994**, *34* (4), 752–781.

(21) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23* (1—3), 3–25.

(22) Borgia, J. A.; Fields, G. B. Chemical synthesis of proteins. *Trends Biotechnol.* **2000**, *18* (6), 243–251.

(23) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29* (2), 97–101.

(24) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11* (23—24), 1046–1053.

(25) Yang, S.-Y. Pharmacophore modeling and applications in drug discovery: challenges and recent advances. *Drug Discovery Today* **2010**, *15* (11—12), 444–450.

(26) Eckert, H.; Bajorath, J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discovery Today* **2007**, *12* (5—6), 225–233.

(27) Warren, G. L.; Andrews, W.; Capelli, A. M.; Clarke, B. P.; LaLonde, J. M.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Peishoff, C. E.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Head, M. S. Critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006**, *49* (20), 5912–5931.