Bioinformatics, 33(17), 2017, 2753–2755
doi: 10.1093/bioinformatics/btx285
Advance Access Publication Date: 4 May 2017
Applications Note

Sequence analysis

modlAMP: Python for antimicrobial peptides

Alex T. Müller, Gisela Gabernet, Jan A. Hiss and Gisbert Schneider*

Department of Chemistry and Applied Biosciences, Swiss Federal Institute of Technology (ETH), CH-8093 Zurich, Switzerland

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on December 30, 2016; revised on April 10, 2017; editorial decision on April 22, 2017; accepted on April 22, 2017

Abstract

Summary: We have implemented the **mo**lecular **d**esign **l**aboratory's **anti**m**i**crobial **p**eptides package (**modlAMP**), a Python-based software package for the design, classification and visual representation of peptide data. modlAMP offers functions for molecular descriptor calculation and the retrieval of amino acid sequences from public or local sequence databases, and provides instant access to precompiled datasets for machine learning. The package also contains methods for the analysis and representation of circular dichroism spectra.

Availability and Implementation: The modlAMP Python package is available under the BSD license from URL http://doi.org/10.5905/ethz-1007-72 or via `pip` from the Python Package Index (PyPI).

Contact: gisbert.schneider@pharma.ethz.ch

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The interest in membranolytic antimicrobial peptides (AMPs) has constantly increased over the last decade (Fjell *et al.*, 2012). Research foci have shifted from isolating natural AMPs towards the computer-assisted design of synthetic analogues and mimetics with improved properties (Jenssen *et al.*, 2008; Juretić *et al.*, 2017). Several successful examples of computationally *de novo* generated AMPs have been reported (Maccari *et al.*, 2013; Müller *et al.*, 2016), together with new online AMP prediction tools (Waghu *et al.*, 2014; Wang *et al.*, 2016). However, to this point one had to connect descriptor calculation, activity prediction and analysis tools through custom scripts, which requires skills in different programming languages and environments. We here present modlAMP, a Python package to ease the discovery and design of novel synthetic AMPs via the amalgamation of sequence generation, descriptor calculation, machine learning and data analysis into a single programming environment. modlAMP provides functions for calculating a variety of different molecular properties and amino acid residue-based peptide descriptors. Furthermore, it enables the *in silico* generation of bespoke peptide libraries with desired properties. The package design follows a modular, object-oriented architecture. The functions used by most of the methods are located in the core module and accessed by other modules through local import. We deliberately kept the number of objects small, and relied on `numpy` (van der Walt *et al.*, 2011) arrays and `pandas` (McKinney, 2010) data frames, where possible. We implemented unit testing to ensure high code quality. The package comes with detailed online documentation (URL https://pythonhosted.org/modlamp), including elaborate examples, that demonstrate the use of the various data classes and analysis methods. A sample script showcases a machine learning workflow for classifying AMPs versus other peptides.

2 Package description

The modlAMP package currently consists of nine modules:

1. `modlamp.descriptors` – molecular descriptor calculations
2. `modlamp.sequences` – *in silico* sequence design
3. `modlamp.database` – queries to peptide databases
4. `modlamp.datasets` – precompiled classification datasets
5. `modlamp.plot` – visualization tools
6. `modlamp.ml` – machine learning models and functions
7. `modlamp.wetlab` – interpretation of experimental data
8. `modlamp.analysis` – comparison of sequence libraries
9. `modlamp.core` – helper functions and parent classes

2.1 Descriptor calculation

The two main classes provided by the `descriptors` module are `GlobalDescriptor` and `PeptideDescriptor`. The available

Table 1. Amino acid property scales available for descriptor calculation through Moreau-Broto type correlation with the `PeptideDescriptor` class

Code-ID[a]	Description	Reference
AASI	Amino acid selectivity index	Juretić *et al.* (2009)
Argos	Argos hydrophobicity scale	Argos *et al.* (2005)
Bulkiness	Side chain bulkiness	Zimmerman *et al.* (1968)
Charge_Phys	Residue charge (pH 7)	Cock *et al.* (2009)
Charge_Acid	Residue charge ($pH < 6$; $H = +1$)	Cock *et al.* (2009)
Eisenberg	Eisenberg consensus scale	Eisenberg *et al.* (1982)
Ez	Energy of lipid bilayer insertion	Senes *et al.* (2007)
Flexibility	Side chain flexibility	Bhaskaran and Ponnuswamy (1988)
GRAVY	Hydrophobicity	Kyte and Doolittle (1982)
Hopp-Woods	Hydrophobicity	Hopp and Woods (1981)
ISA–ECI	Isotropic surface area–electronic charge index	Collantes and Dunn (1995)
Janin	Hydrophobicity	Cornette *et al.* (1987)
KyteDoolittle	Hydrophobicity	Kyte and Doolittle (1982)
Levitt_alpha	α-helical propensity	Levitt (1978)
MSS	Side chain topological shape and size	Raychaudhury *et al.* (1999)
MSW	Principal components of steric and 3D residue properties	Zaliani and Gancia (1999)
pepCATS	Binary pharmacophoric features	Koch *et al.* (2013)
Polarity	Amino acid polarity	Zimmerman *et al.* (1968)
PPCALI	Principal components of selected side chain properties	Koch *et al.* (2013)
Refractivity	Relative refractivity values	McMeekin *et al.* (1962)
t_scale	Principal components of GRID derived values	Cocchi and Johansson (1993)
TM_tend	Transmembrane propensity	Zhao and London (2006)
z3	Original three-dimensional Z-scale	Hellberg *et al.* (1987)
z5	Extended five-dimensional Z-scale	Sandberg *et al.* (1998)

Optionally, users can use their own, locally saved amino acid property scales.

[a] Code-ID refers to the `scalename` input option of the `PeptideDescriptor` class.

property scales and the corresponding `scalename` option codes for `PeptideDescriptor` instantiation are listed in Table 1.

Holistic, one-dimensional peptide representations, e.g. total charge, molecular weight, hydrophobic ratio, or aromaticity, are calculated in the `GlobalDescriptor` class. The `PeptideDescriptor` class handles property-based descriptors computed by Moreau-Broto correlation functions with variable sliding sequence windows (Broto *et al.*, 1984). Amino acid sequences can be imported as individual residue strings, a list of strings, or in FASTA format, *e.g.* as follows:

```
⋙ from modlamp.descriptors import PeptideDescriptor
⋙ desc = PeptideDescriptor(‘GLFDIVKKVVGALGSL’,
‘pepCATS’)
⋙ desc.calculate_crosscorr(window=7)
⋙ desc.descriptor
array([[0.6875, 0.46666667, 0.42857143, ...]])
```

2.2 Sequence generator

The `sequence` module implements ten classes for *in silico* sequence generation (Supplementary Table S1), namely (i) random sequences, (ii) sequences with a presumed amphipathic helical structure, (iii) kinked amphipathic helices, (iv) amphipathic helices with a defined hydrophobic arc, (v) sequences with a linear hydrophobicity gradient, (vi) centrosymmetric sequences, (vii) sequences incorporating a possible heparin binding domain, (viii) sequences generated from frequent AMP *n*-grams, (ix) sequences with the residue probability of known helical anticancer peptides and (x) mixed peptide libraries.

2.3 Visualization

The `plot` module contains several functions for data visualization from the `matplotlib` package (URL https://matplotlib.org) (Fig. 1). In addition, `GlobalAnalysis` from the analysis module provides a graphical overview of the properties of given sequence libraries (Supplementary Fig. S1).

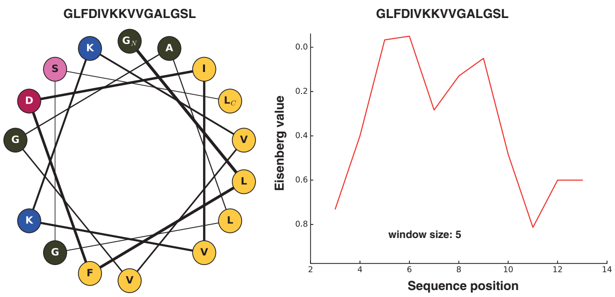

Fig. 1. Examples of helical wheel (*left*) and hydrophobicity (*right*) plots generated with the `helical_wheel` and `plot_profile` functions

2.4 Machine learning

modlAMP provides standard functions for machine learning and model selection via a pipeline of data scaling, parameter grid search and model cross-validation for both support vector machine (Cortes and Vapnik, 1995) and random forest classifiers (Breiman, 2001). For example, the function `ml.train_best_model` performs a parameter grid search on the selected model and training dataset. As the name implies, it returns the best performing model based on the Matthews correlation coefficient (Matthews, 1975) obtained by cross-validation. The functions `ml.score_cv` and `ml.score_testset` evaluate the performance of existing classifiers by performing cross-validation or calculating the test set error, respectively. The function `ml.predict` retrieves the *pseudo*-probability of custom-generated peptides to belong to the different classes, and thereby informs the user about the model's estimated uncertainty and applicability domain. modlAMP relies on the `scikit-learn` package (Pedregosa *et al.*, 2011), providing thoroughly tested state-of-the-art implementations of machine learning and data preprocessing methods in Python.

2.5 Circular dichroism spectral analysis

Secondary structure dynamics may be a major feature determining antimicrobial activity of certain classes of AMPs. Initial laboratory experiments usually include circular dichroism (CD) spectroscopy of peptides in different solvents. modlAMP contains the `wetlab` module for the analysis of CD data (Supplementary Fig. S2) and signal transformation to mean residue ellipticity.

3 Conclusions

The modlAMP package provides an application programming interface that efficiently facilitates the handling of large sets of peptide sequences. It gives access to a full pipeline of *in silico* methods for peptide analysis and design, ranging from molecular descriptor calculation to machine learning. The software is provided as open source under the BSD-3 license (URL https://opensource.org/licenses/BSD-3-Clause) from the Python Package Index (URL https://pypi.python.org/pypi/modlamp) and can be installed through pip (`pip install modlamp`). Full documentation of the package, including use cases and sample applications, together with a detailed explanation of all classes and functions, is available from URL https://pythonhosted.org/modlamp/.

Acknowledgements

The authors thank Arndt R. Finkelmann for technical support.

Funding

This research was financially supported by the Swiss National Science Foundation (grants no. 200021_157190, CRSII2_160699).

Conflict of Interest: G. S. is a co-founder of inSili.com LLC, Zurich.

References

Argos,P. *et al.* (2005) Structural prediction of membrane-bound proteins. *Eur. J. Biochem.*, **128**, 565–575.

Bhaskaran,R. and Ponnuswamy,P.K. (1988) Positional flexibilities of amino acid residues in globular proteins. *Int. J. Pept. Protein Res.*, **32**, 241–255.

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Broto,P. *et al.* (1984) Molecular structures: perception, autocorrelation descriptor and SAR studies: system of atomic contributions for the calculation of the n-octanol/water partition coefficients. *Eur. J. Med. Chem.*, **19**, 71–78.

Cocchi,M. and Johansson,E. (1993) Amino acids characterization by GRID and multivariate data analysis. *Quant. Struct. Act. Relation.*, **12**, 1–8.

Cock,P.J.A. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.

Collantes,E.R. and Dunn,W.J. (1995) Amino acid side chain descriptors for quantitative structure-activity relationship studies of peptide analogs. *J. Med. Chem.*, **38**, 2705–2713.

Cornette,J.L. *et al.* (1987) Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol.*, **195**, 659–685.

Cortes,C. and Vapnik,V. (1995) Support-vector networks. *Mach. Learn.*, **20**, 273–297.

Eisenberg,D. *et al.* (1982) The helical hydrophobic moment: a measure of the amphiphilicity of a helix. *Nature*, **299**, 371–374.

Fjell,C.D. *et al.* (2012) Designing antimicrobial peptides: form follows function. *Nat. Rev. Drug Discov.*, **11**, 37–51.

Hellberg,S. *et al.* (1987) Peptide quantitative structure-activity relationships, a multivariate approach. *J. Med. Chem.*, **30**, 1126–1135.

Hopp,T.P. and Woods,K.R. (1981) Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci.*, **78**, 3824–3828.

Jenssen,H. *et al.* (2008) QSAR modeling and computer-aided design of antimicrobial peptides. *J. Pept. Sci.*, **14**, 110–114.

Juretić,D. *et al.* (2009) Computational design of highly selective antimicrobial peptides. *J. Chem. Inf. Model.*, **49**, 2873–2882.

Juretić,D. *et al.* (2017) Tools for designing amphipathic helical antimicrobial peptides. *Methods Mol. Biol.*, **1548**, 23–34.

Koch,C.P. *et al.* (2013) Scrutinizing MHC-I binding peptides and their limits of variation. *PLoS Comput. Biol.*, **9**, e1003088.

Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.

Levitt,M. (1978) Conformational preferences of amino acids in globular proteins. *Biochemistry*, **17**, 4277–4285.

Maccari,G. *et al.* (2013) Antimicrobial peptides design by evolutionary multiobjective optimization. *PLoS Comput. Biol.*, **9**, e1003212.

Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.

McKinney,W. (2010) Data structures for statistical computing in Python. *Proc. 9th Python Sci. Conf.*, **1697900**, 51–56.

McMeekin,T.L. *et al.* (1962) Refractive indices of proteins in relation to amino acid composition and specific volume. *Biochem. Biophys. Res. Commun.*, **7**, 151–156.

Müller,A.T. *et al.* (2016) Sparse Neural Network Models of Antimicrobial Peptide-Activity Relationships. *Mol. Inf.*, **35**, 606–614.

Pedregosa,F. *et al.* (2011) Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Raychaudhury,C. *et al.* (1999) Topological shape and size of peptides: identification of potential allele specific helper T cell antigenic sites. *J. Chem. Inf. Comput. Sci.*, **39**, 248–254.

Sandberg,M. *et al.* (1998) New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *J. Med. Chem.*, **41**, 2481–2491.

Senes,A. *et al.* (2007) Ez, a depth-dependent potential for assessing the energies of insertion of amino acid side-chains into membranes: derivation and applications to determining the orientation of transmembrane and interfacial helices. *J. Mol. Biol.*, **366**, 436–448.

Waghu,F.H. *et al.* (2014) CAMP: Collection of sequences and structures of antimicrobial peptides. *Nucleic Acids Res.*, **42**, 1154–1158.

van der Walt,S. *et al.* (2011) The NumPy array: a structure for efficient numerical computation. *Comput. Sci. Eng.*, **13**, 22–30.

Wang,G. *et al.* (2016) APD3: the antimicrobial peptide database as a tool for research and education. *Nucleic Acids Res.*, **44**, D1087–D1093.

Zaliani,A. and Gancia,E. (1999) MS-WHIM scores for amino acids: a new 3D-description for peptide QSAR and QSPR studies. *J. Chem. Inf. Comput. Sci.*, **39**, 525–533.

Zhao,G. and London,E. (2006) An amino acid 'transmembrane tendency' scale that approaches the theoretical limit to accuracy for prediction of transmembrane helices: relationship to biological hydrophobicity. *Protein Sci.*, **15**, 1987–2001.

Zimmerman,J.M. *et al.* (1968) The characterization of amino acid sequences in proteins by statistical methods. *J. Theor. Biol.*, **21**, 170–201.