Contents lists available at ScienceDirect

Data in brief

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



Data Article

# Data of small peptides in SMILES and three-dimensional formats for virtual screening campaigns



## Vivitri Dewi Prasasty<sup>a</sup>, Enade Perdana Istyastono<sup>b,\*</sup>

<sup>a</sup> Faculty of Biotechnology, Atma Jaya Catholic University of Indonesia, Jakarta, 12930, Indonesia
<sup>b</sup> Faculty of Pharmacy, Sanata Dharma University, Paingan, Maguwoharjo, Sleman, Depok, Yogyakarta, 55282, Indonesia

## ARTICLE INFO

Article history: Received 3 April 2019 Received in revised form 19 September 2019 Accepted 26 September 2019 Available online 4 October 2019

Keywords: Small peptide Drug discovery Virtual screening Dipeptide Tripeptide Tetrapeptide

## ABSTRACT

The data presented in this article are structures of dipeptides, tripeptides and tetrapeptides constructed from all possible combinations of 20 natural and common amino acids. In total, the data contains 168400 peptides. The structures are available in their simplified molecular-input line-entry system (SMILES) and three-dimensional (3D) formats. The type of data are text files, which could be accessed and modified either by text editor applications (e.g. Notepad++) or by molecule visualization softwares (e.g., YASARA View). These structures could be used further in virtual screening campaigns in the early stage of drug discovery projects. © 2019 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

## 1. Data

There are three files in the data: dipeptides.zip, tripeptides.zip, and tetrapeptides.zip. Each file contains two subdirectories, *i.e.*, 3d-in-mol2 and smi. The directories 3d-in-mol2 and smi contain structures in 3D in mol2 format and structures in SMILES format, respectively. For dipeptides.zip and

\* Corresponding author.

https://doi.org/10.1016/j.dib.2019.104607

E-mail address: enade@usd.ac.id (E.P. Istyastono).

<sup>2352-3409/© 2019</sup> The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http:// creativecommons.org/licenses/by/4.0/).

#### Specifications Table

Subject area	Pharmaceutical Sciences
More specific subject area	Drug discovery
Type of data	Text file
How data was acquired	Molecular modelling
Data format	Raw
Experimental factors	The data were produced by employing a computer server with Ubuntu Linux 14.04.5 as the operating system, 8GB RAM and 8 virtual CPUs of 6.4 GHz
Experimental features	The data were built by employing two main computational chemistry tools, i.e., molconvert 17.13.0 from ChemAxon and gen3d from Open Babel version 2.3.0.
Data source location	Data were built using molecular modelling techniques in a virtual laboratory owned by Sanata Dharma University, Yogyakarta, Indonesia.
Data accessibility	Mendeley Data. doi:http://dx.doi.org/10.17632/z8zh5rpthg.1
Related research article	Prasasty, V. Radifar, M., Istyastono, E., 2018. Natural peptides in drug discovery targeting acetylcholinesterase. Molecules. 23(9): 2344. https://doi.org/10.3390/molecules23092344 [1]

#### Value of the Data

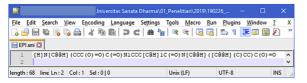
- The data provide structures of all possible natural dipeptides, tripeptides and tetrapeptides to be employed further in virtual screening campaigns.
- The SMILES formats of the peptides provide opportunities to perform virtual screening campaigns from the very beginning starting point.
- The 3D structures in mol2 formats provide opportunities to perform virtual screening campaigns from the 3D structure as the starting point.
- Since most of the peptides have molecular weight of less than 500 Da, the resulted hits in the virtual screening campaigns could be optimized further to be potential ligands in drug discovery projects.

tripeptides, those directories contain structure files of the peptides named by their sequences in oneletter-code. These dipeptides and tripeptides have been subjected in a structure-based virtual screening to discover novel acetylcholinesterase (AChE) inhibitors [1] by employing a retrospectively validated protocol [2]. In tetrateptides.zip, the 3d-in-mol2 and smi directories contain subdirectories named with the first sequence of the peptides inside the directory followed by triple x. For example, Fig. 1 shows tripeptide EPI in the smi and the mol2 formats visualized using both Notepad++ as a text editor and Yasara View as a molecule visualization application. The file EPI.smi could be found in the subdirectory smi, while the file EPI.mol2 could be found in the subdirectory 3d-in-mol2 in tripeptides.zip.

### 2. Experimental design, materials, and methods

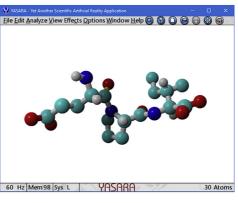
The data were built in a computer server with Ubuntu Linux 14.04.5 as the operating system, 8GB RAM and 8 virtual CPUs of 6.4 GHz. First of all, three input files named dipeptides.lst, tripeptides.lst, and tetrapeptides.lst were made. These files contained all possible sequences of dipeptides, tripeptides, and tetrapeptides in their one-letter-code, respectively. Subsequently, the smi files were built by using module *molconvert* (Molecule File Converter, version 17.13.0) from ChemAxon (https://chemaxon.com/). Based on these smi files, the 3D structures were then built using module *gen3d* from Open Babel version 2.3.0 [3]. The *molconvert* module requires amino acids sequence to be converted to SMILES format. The following is, for example, the generic code to build the dipeptides presented in this article: for i in \$(cat dipeptides.lst); do mkdir \$i; cd \$i; molconvert –peptide \$i smiles > \$i.smi; babel –title \$i –gen3d -ismi \$i.smi -omol2 \$i.mol2; cd ...; done.

The availability of the data will significantly reduce time to perform virtual screening campaigns. The following are some examples of study that could benefit from the data: (i) PeptoGrid - Rescoring function for AutoDock Vina to identify new bioactive molecules from short peptide libraries [4],



(A)

2	🗎 🛍 🗟 🐚 🚔	% <b>b b þ</b>	C 🗎 🛗 🎭	👒 👒   [	3 🗟	Ep 11	1 🖉 🖉	J 🔊	>>	
EPI.s	mi 🖂 🔚 EPI.mol2 🔛									
1	@ <tripos>MOLECU</tripos>	LE							^	
2	EPI									
3	52 52 0 0 0									60 Hz Mem 98
4	SMALL									
5	GASTEIGER									
6										
7	@ <tripos>ATOM</tripos>									
8	1 H1	9.5193	0.4105	-0.005		1	GLU1			
9	2 N	8.6761	-0.1528	0.165		1	GLU1		·	VASARA - Yet Anothe
10	3 CA	7.6353	0.7314	0.728		1	GLU1			File Edit Analyze V
11	4 HA	8.0549	1.2843	1.576		1	GLU1			
12	5 CB	7.1965	1.7862	-0.304		1	GLU1			
13	6 CG	6.6911	1.2930	-1.662		1	GLU1			
14	7 CD	7.7677	0.9502	-2.665		1	GLU1			
15	8 OE2	8.4509	-0.1776	-2.389		1	GLU1			
16	9 OE1	7.9845	1.6099	-3.666		1	GLU1			
17	10 C	6.4634	-0.1423	1.249		1	GLU1			
18	11 0	5.5006	-0.4356	0.537		1	GLU1			
19	12 N	6.5399	-0.6331	2.538			PRO2			
20	13 CD	7.6466	-0.4911	3.470		2	PRO2			
21	14 CG	7.6803	-1.8237	4.193		2	PRO2 PRO2			
22	15 CB 16 CA	6.2087 5.5546	-2.1930	4.312		2	PRO2 PRO2			
23 24	16 CA 17 HA	5.4088	-2.3805			2	PRO2 PRO2			
24	17 HA 18 C	4.1629	-2.3805	2.296		2	PRO2 PRO2			
25	19 0	4.1629	-0.9850	4.443		2	PRO2 PRO2			
20	20 N	3.5287	-0.3895	2.302			ILE3			
28	20 N 21 CA	2.1767	0.2059	2.386		3	ILE3			
29	22 HA	2.0244	0.4944	3.434		3	ILES			
30	23 CB	2.0244	1.4768	1.492		3	ILES			
31	24 HB	2.8563	2.1656	1.827		3	ILES		~	
:	24 115	2.0000	2.1050	1.027.		, in the second se	THES	>		
ngth : S	5,332   Ln:1 Col:1	Sel:010		Unix (LF)		UTF-8		INS		60 Hz Mem 98



(B)

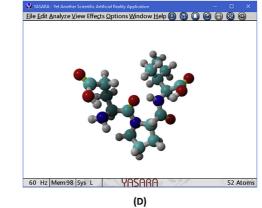


Fig. 1. The EPI.smi in Notepad++ (A) and YASARA View (B), and the EPI.mol2 in Notepad++ (C) and YASARA View (D).

(ii) Structure-based virtual screening for fragment-like ligands of the G protein-coupled histamine H4 receptor [5], and (iii) Natural peptides in drug discovery targeting acetylcholinesterase [1].

## Acknowledgments

This work was supported by the Ministry of Research, Technology and Higher Education of the Republic of Indonesia [grant number: 17/AKM/PNT/2019].

## **Conflict of Interest**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

- V. Prasasty, M. Radifar, E. Istyastono, Natural peptides in drug discovery targeting acetylcholinesterase, Molecules 23 (2018) 2344, https://doi.org/10.3390/molecules23092344.
- [2] F.D.O. Riswanto, M. Hariono, S.H. Yuliani, E.P. Istyastono, Computer-aided design of chalcone derivatives as lead compounds targeting acetylcholinesterase, Indones. J. Pharm. 28 (2017) 100–111. http://indonesianjpharm.farmasi.ugm.ac.id/index. php/3/article/view/1182.
- [3] N.M. O'Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, G.R. Hutchison, Open Babel: an open chemical toolbox, J. Cheminf. 3 (2011) 33–47, https://doi.org/10.1186/1758-2946-3-33.
- [4] A.O. Zalevsky, A.S. Zlobin, V.R. Gedzun, R.V. Reshetnikov, M.L. Lovat, A.V. Malyshev, I.I. Doronin, G.A. Babkin, A.V. Golovin, Peptogrid—rescoring function for autodock vina to identify new bioactive molecules from short peptide libraries, Molecules 24 (2019) 277, https://doi.org/10.3390/molecules24020277.
- [5] E.P. Istyastono, A.J. Kooistra, H. Vischer, M. Kuijer, L. Roumen, S. Nijmeijer, R. Smits, I. de Esch, R. Leurs, C. de Graaf, Structurebased virtual screening for fragment-like ligands of the G protein-coupled histamine H4 receptor, Med. Chem. Commun. 6 (2015) 1003–1017, https://doi.org/10.1039/C5MD00022J.