SeqDivA: Sequence Diversity Analysis Tool for Detecting the Twilight Zone of Alignment Algorithms

Guillermin Agüero-Chapin[1,2]* and Evys Ancede-Gallardo[3]

[1] CIIMAR/CIMAR, Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Terminal de Cruzeiros do Porto de Leixões, Av. General Norton de Matos s/n 4450-208 Matosinhos, Porto, Portugal.

[2] Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, 4169-007 Porto, Portugal

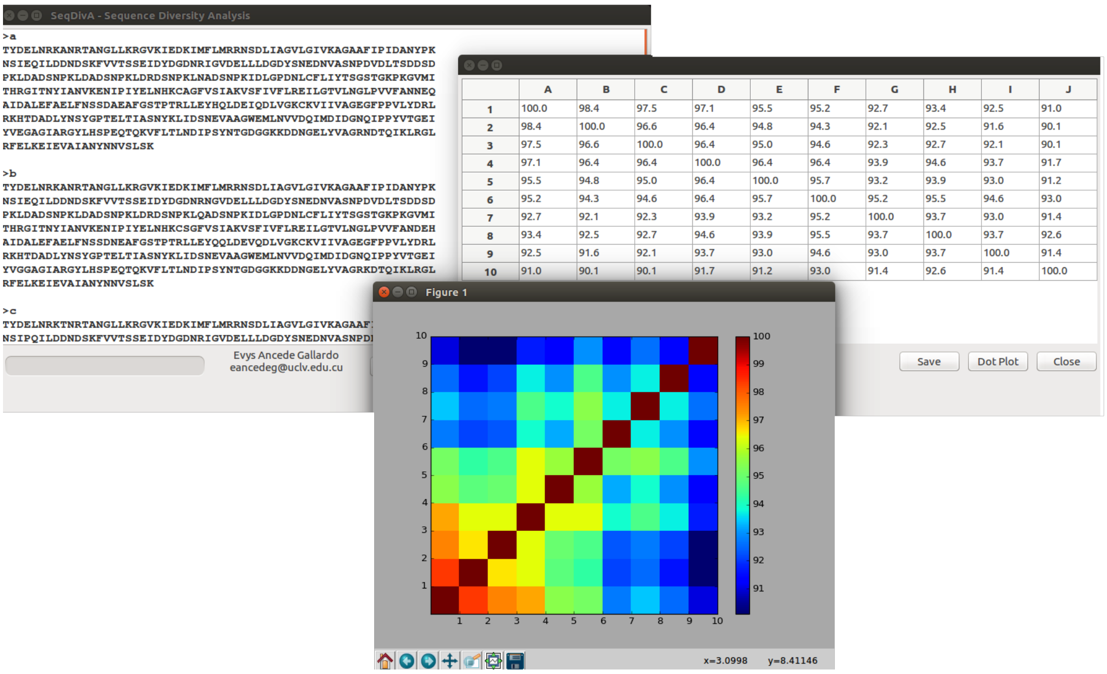
[3] Facultad de Ciencias Exactas, Universidad Andrés Bello, República 275, Santiago, Chile

*Corresponding author: GACh: gchapin@ciimar.up.pt

Abstract

Looking into the literature and scientific forums, there isn't any software that can explore the diversity of a database or a sequence subset by applying the similarity measures reported to delimit the twilight zone according all previously mentioned thresholds. So far, in order to retrieve several similarity measures like identity, similarity and scores in an all-*vs*-all pairwise sequence comparison, users should run previously software like needle (global alignment), water (local alignment), blast (local alignment) and even multiple sequence alignments (MSAs) tools (http://imed.med.ucm.es/Tools/sias.html), then results should be parsed to be presented in a *nxn* matrix. However, going through all these steps to get at the final similarity matrix require programming skills.

Here, we present SeqDivA, a python-based tool with a friendly GUI allowing non-expert users to run alignment algorithms (water, needle and blast) to compare all *vs* all protein, DNA and RNA sequences (**Figure 1**). SeqDivA provides similarity, identity and bit-score matrixes to explore the diversity/homology of the sequences, enabling the delimitation of the twilight zone. The resulting matrixes are visualized using dot plot-like graphs representing pairwise similarity measures (identities, similarity and bit-scores). SeqDivA also allows redundancy reduction by exploring amino acid identities from global alignments and can be connected to the output of software simulating related sequences with a known evolutionary history i.e. ROSE [1] and INDELible [2] in order to get subsets of homologous sequences at different identities or bit-scores ranges. The software can be freely downloaded at https://github.com/eancedeg/SeqDivA. The software was published as part of the paper published at https://doi.org/10.3390/biom10010026.



```
>a
TYDELNRKANRTANGLLKRGVKIEDKIMFLMRRNSDLIAGVLGIVKAGAAFIPIDANYPK
NSIEQILDDNDSKFVVTSSEIDYDGDNRIGVDELLLDGDYSNEDNVASNPDVDLTSDDSD
PKLDADSNPKLDADSNPKLDRDSNPKLNADSNPKIDLGPDNLCFLIYTSGSTGKPKGVMI
THRGITNYIANVKENIPIYELNHKCAGFVSIAKVSFIVFLREILGTVLNGLPVVFANNEQ
AIDALEFAELFNSSDAEAFGSTPTRLLEYHQLDEIQDLVGKCKVIIVAGEGFPPVLYDRL
RKHTDADLYNSYGPTELTIASNYKLIDSNEVAAGWEMLNVVDQIMDIDGNQIPPYVTGEI
YVEGAGIARGYLHSPEQTQKVFLTLNDIPSYNTGDGGKKDDNGELYVAGRNDTQIKLRGL
RFELKEIEVAIANYNNVSLSK

>b
TYDELNRKANRTANGLLKRGVKIEDKIMFLMRRNSDLIAGVLGIVKAGAAFIPIDANYPK
NSIEQILDDNDSKFVVTSSEIDYDGDNRNGVDELLLDGDYSNEDNVASNPDVDLTSDDSD
PKLDADSNPKLDADSNPKLDRDSNPKLQADSNPKIDLGPDNLCFLIYTSGSTGKPKGVMI
THRGITNYIANVKENIPIYELNHKCSGFVSIAKVSFIVFLREILGTVLNGLPVVFANDEH
AIDALEFAELFNSSDNEAFGSTPTRLLEYQQLDEVQDLVGKCKVIIVAGEGFPPVLYDRL
RKHTDADLYNSYGPTELTIASNYKLIDSNEVAAGWEMLNVVDQIMDIDGNQIPPYVTGEI
YVGGAGIARGYLHSPEQTQKVFLTLNDIPSYNTGDGGKKDDNGELYVAGRKDTQIKLRGL
RFELKEIEVAIANYNNVSLSK

>c
TYDELNRKTNRTANGLLKRGVKIEDKIMFLMRRNSDLIAGVLGIVKAGAAF
NSIPQILDDNDSKFVVTSSEIDYDGDNRIGVDELLLDGDYSNEDNVASNPD
```

	A	B	C	D	E	F	G	H	I	J
1	100.0	98.4	97.5	97.1	95.5	95.2	92.7	93.4	92.5	91.0
2	98.4	100.0	96.6	96.4	94.8	94.3	92.1	92.5	91.6	90.1
3	97.5	96.6	100.0	96.4	95.0	94.6	92.3	92.7	92.1	90.1
4	97.1	96.4	96.4	100.0	96.4	96.4	93.9	94.6	93.7	91.7
5	95.5	94.8	95.0	96.4	100.0	95.7	93.2	93.9	93.0	91.2
6	95.2	94.3	94.6	96.4	95.7	100.0	95.2	95.5	94.6	93.0
7	92.7	92.1	92.3	93.9	93.2	95.2	100.0	93.7	93.0	91.4
8	93.4	92.5	92.7	94.6	93.9	95.5	93.7	100.0	93.7	92.6
9	92.5	91.6	92.1	93.7	93.0	94.6	93.0	93.7	100.0	91.4
10	91.0	90.1	90.1	91.7	91.2	93.0	91.4	92.6	91.4	100.0

Figure 1. Screen shot of the SeqDivA's GUI. The input fasta file made up by 10 hypothetical protein sequences and the main outputs: the identity matrix all-vs-all and the dotplot representing the identity/similarly/bitscore variation among the sequence pairs.

References

1- Stoye, J., D. Evers, and F. Meyer, *Rose: generating sequence families.* Bioinformatics, 1998. **14**(2): p. 157-163.

2- Fletcher, W. and Z. Yang, *INDELible: a flexible simulator of biological sequence evolution.* Mol Biol Evol, 2009. **26**(8): p. 1879-88