EP-pred: a machine learning tool for bioprospecting promiscuous ester hydrolases

Ruite Xiang1,†, Laura Fernandez-Lopez2,†, Ana Robles1, Manuel Ferrer2 and Victor Guallar1,3\*

1 Department of Life Sciences, Barcelona Supercomputing Center (BSC), 08034 Barcelona, Spain; [ruite.xiang@bsc.es](mailto:ruite.xiang@bsc.es), [ana.robles@bsc.es](mailto:ana.robles@bsc.es), [victor.guallar@bsc.es](mailto:victor.guallar@bsc.es)

2 Department of Applied Biocatalysis, ICP, CSIC, 28049 Madrid, Spain; [l.fernandez.lopez@csic.es](mailto:l.fernandez.lopez@csic.es), [mferrer@icp.csic.es](mailto:mferrer@icp.csic.es)

3 Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain, 08010 Barcelona, Spain; [victor.guallar@bsc.es](mailto:victor.guallar@bsc.es)

**Supplementary information**

|  |
| --- |
|  |

**Ifeature features**

|  |  |
| --- | --- |
| **iFeaure descriptors** | **definition** |
| Grouped amminoacid composition | aa composition and its variants are the % of each aa, dipeptide, tripeptide in the sequence. |
| Grouped dipeptide, tripeptide composition |
| K-spaced aa group pairs composition |
| Moran Autocorrelation | It describes the correlation between 2 aa in terms of 8 aa properties |
| Geary Autocorrelation |
| Normalized Moreau-Broto Autocorrelation |
| Composition | The 20 aas are divided into 3 groups for each of the 7 properties considered and then it looks the class composition, how frequent are the transitions to other classes and the overall distribution along the sequences. |
| Transition |
| Distribution |
| Conjoint triad | It clusters 20 aa into 7 classes, it treats tripeptides as an unit and counts the frequency of each unit in the sequence. |
| K-spaced conjoint triad |
| Sequence-order coupling number | It can be used to represent aa distribution patterns of a specific physicochemical property along a protein |
| Quasi sequence order |
| Pseudo-aa-composition | It counts the composition of the aa in the sequence, but it adds a coupling term to not lose order information of the aa. |
| Amphiphilic Pseudo aa composition |
| Pseudo K-tuple reduced aa composition type1 to 16 |

**Table S1** | The features extracted from iFeature [1] used in this project

**Possum Features**

**Table S2** | The features extracted from possum [2]

|  |  |
| --- | --- |
| **Mathematical transformations applied** | **Possum descriptors** |
| Row transformations | AAC-PSSM |
| D-FPSSM |
| S-FPSSM |
| AB-PSSM |
| RPM-PSSM |
| smoothed-PSSM window size (5) |
| smoothed-PSSM window size (7) |
| smoothed-PSSM window size (9) |
| PSSM-composition |
| Column transformations | DPC-PSSM |
| k-separated-bigrams-PSSM |
| tri-gram-PSSM |
| EEDP |
| TPC |
| Mixed of row and column transformations | EDP |
| Pse-PSSM (ξ = 1) |
| Pse-PSSM (ξ = 2) |
| Pse-PSSM (ξ = 3) |
| DP-PSSM |
| PSSM-AC |
| PSSM-CC |
| RPSSM |

[1] Z. Chen *et al.*, ‘iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences’, *Bioinformatics*, vol. 34, no. 14, pp. 2499–2502, Jul. 2018, doi: 10.1093/bioinformatics/bty140.

[2] J. Wang *et al.*, ‘POSSUM: a bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles’, *Bioinformatics*, vol. 33, no. 17, pp. 2756–2758, Sep. 2017, doi: 10.1093/bioinformatics/btx302.