Supplementary information for:

pICalculax: Improved Prediction of Isoelectric Point for Modified Peptides

Esben J. Bjerrum[1;2*], Jan H. Jensen[2] and Jakob L. Tolborg[3]

1. Wildcard Pharmaceutical Consulting, Frødings Alle 41, 2860 Søborg, Denmark
2. Biochemfusion Aps, Løvspringsvej 4C, 1.tv, 2920 Charlottenlund, Denmark
3. Zealand Pharma A/S, Smedeland 36, 2600 Glostrup, Denmark
* Corresponding Author: esben@wildcardconsulting.dk

Abstract

The isoelectric point of a peptide is a physico-chemical property that can be accurately predicted from the sequence of the peptide when the peptide is built from natural amino acids only. Peptides can however have chemical modifications, such as phosphorylations and amidations, which can result in erroneous predictions if not accounted for. The use of unnatural amino acids in peptide drug discovery can also exhibit an influence on the pI. Here we report on an open source program, pICalculax, which in an extensible way can handle pI calculations of modified peptides. Test on a database of modified peptides and experimentally determined pI values show an improvement in pI predictions when taking the modifications into account. The correlation coefficient improves from 0.45 to 0.91, and the root mean square deviation likewise improves from 3.3 to 0.9. The program is available at https://github.com/EBjerrum/pICalculax

Supplementary Table 1: Comparison of average performance on modified and unmodified datasets for pIR and pICalculax. Average correlation coefficient (R^2) and average root mean square deviation (RMSD) are for all tested pKa sets and for the modified and unmodified datasets as FASTA sequences or protein line notation (PLN). For pIR the PLN was converted to FASTA with markup of the possible modifications known by pIR.

Program / Datasets	**Avg. R^2**		**Avg. RMSD**	
	FASTA	**PLN**	**FASTA**	**PLN**
pICalculax				
Unmodified peptides	0.89	0.90	0.66	0.68
Modified peptides	0.48	0.86	2.57	0.68
pIR				
Unmodified peptides	0.89	0.89	0.66	0.66
Modified Peptides	0.48	0.65	2.57	1.75

Supplementary Table 2: All benchmark results for the unmodified datasets.

		pICalculax				pIR			
		FASTA		PLN		FASTA		PLN	
		R2	RMSE	R2	RMSE	R2	RMSE	R2	RMSE
EMBOSS	**Gauci**	0.86	0.45	0.85	0.47	0.84	0.46	0.84	0.46
	Heller	0.92	0.59	0.93	0.62	0.92	0.59	0.92	0.59
	Reaxys	0.99	0.66	0.99	0.66	0.99	0.66	0.99	0.66
Grimsley	**Gauci**	0.90	0.32	0.91	0.33	0.90	0.32	0.90	0.32
	Heller	0.90	0.61	0.94	0.49	0.90	0.61	0.90	0.61
	Reaxys	0.99	0.38	0.99	0.38	0.99	0.38	0.99	0.38
IPC	**Gauci**	0.76	0.60	0.74	0.66	0.76	0.60	0.76	0.60
	Heller	0.91	0.77	0.92	0.83	0.91	0.77	0.91	0.77
	Reaxys	0.97	0.88	0.97	0.88	0.97	0.88	0.97	0.88
Lehninger	**Gauci**	0.75	0.62	0.74	0.66	0.75	0.62	0.75	0.62
	Heller	0.90	0.79	0.92	0.84	0.90	0.79	0.90	0.79
	Reaxys	0.96	0.91	0.96	0.91	0.96	0.90	0.96	0.90
Solomon	**Gauci**	0.75	0.61	0.73	0.67	0.75	0.61	0.75	0.61
	Heller	0.90	0.78	0.92	0.84	0.90	0.78	0.90	0.78
	Reaxys	0.96	0.89	0.96	0.89	0.96	0.89	0.96	0.89
Average:		**0.89**	**0.66**	**0.90**	**0.68**	**0.89**	**0.66**	**0.89**	**0.66**

Supplementary Table 3: All benchmark results for the modified datasets.

		pICalculax				pIR			
		FASTA		PLN		FASTA		PLN	
		R2	RMSE	R2	RMSE	R2	RMSE	R2	RMSE
EMBOSS	Gauci	0.50	2.15	0.82	0.40	0.51	2.13	0.83	0.41
	Reaxys	0.45	2.92	0.92	0.76	0.45	2.93	0.45	2.93
Grimsley	Gauci	0.53	1.83	0.85	0.55	0.53	1.83	0.86	0.53
	Reaxys	0.45	3.30	0.92	0.89	0.45	3.30	0.45	3.30
IPC	Gauci	0.45	2.34	0.79	0.62	0.45	2.34	0.80	0.61
	Reaxys	0.50	2.84	0.91	0.77	0.50	2.84	0.50	2.84
Lehninger	Gauci	0.45	2.35	0.79	0.65	0.45	2.35	0.80	0.64
	Reaxys	0.51	2.82	0.90	0.81	0.51	2.82	0.51	2.82
Solomon	Gauci	0.45	2.34	0.79	0.62	0.45	2.34	0.80	0.61
	Reaxys	0.50	2.83	0.91	0.77	0.50	2.83	0.50	2.83
Average:		**0.48**	**2.57**	**0.86**	**0.68**	**0.48**	**2.57**	**0.65**	**1.75**

Supplementary File: atomic_data.cpp.patch

Patch file for atom_constants.cpp from rdkit (https://github.com/rdkit/rdkit), which allow support for pseudo atoms. Patch-file generated based on RDKit release 2015_09_2.

Supplementary File: proteax_desktop_mod_db_export.sdf

SD file export of the modification database for Proteax Desktop (http://www.biochemfusion.com/downloads/#ProteaxDesktop).

can be loaded with the ProteaxAdmin.exe client. Linux users can run the client using wine (https://www.winehq.org/)

Supplementary Spreadsheet: Reaxys_predicted.xls

Excel files with test database of peptides in PLN notation, literature pI values and predicted pI values.