Journal of
CHEMICAL INFORMATION
AND MODELING

Application Note

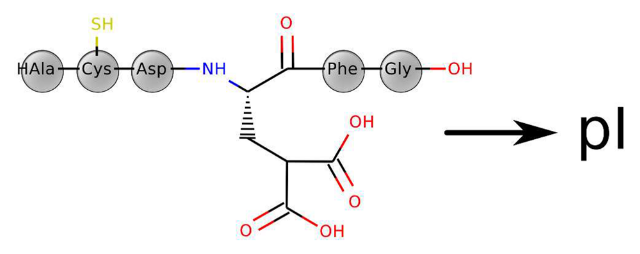
pubs.acs.org/jcim

pICalculax: Improved Prediction of Isoelectric Point for Modified Peptides

Esben J. Bjerrum,[*,†,‡] Jan H. Jensen,[‡] and Jakob L. Tolborg[§]

[†]Wildcard Pharmaceutical Consulting, Frødings Alle 41, 2860 Søborg, Denmark
[‡]Biochemfusion Aps, Løvspringsvej 4C, 1.tv, 2920 Charlottenlund, Denmark
[§]Zealand Pharma A/S, Smedeland 36, 2600 Glostrup, Denmark

Supporting Information



ABSTRACT: The isoelectric point of a peptide is a physicochemical property that can be accurately predicted from the sequence of the peptide when the peptide is built from natural amino acids. Peptides can however have chemical modifications, such as phosphorylations, amidations, and unnatural amino acids, which can result in erroneous predictions if not accounted for. Here we report on an open source program, pICalculax, which in an extensible way can handle pI calculations of modified peptides. Tests 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

INTRODUCTION

The isoelectric point (pI) of a protein or peptide is the pH at which the net charge on the molecular ensemble is zero. This physicochemical property can be determined experimentally with experiments such as fixed pH gradient gel electrophoresis[1] or capillary electrophoresis.[2] When the pH of the solution matches the pI, the solute will not migrate in an electric field. Knowledge about the pI can lead to a better understanding of function[3] and be of practical aid in the laboratory to avoid precipitation and understand behavior during ion exchange chromatography.

The pI of a regular peptide under denaturing conditions can be predicted from the sequence.[4] By counting the acidic and basic groups and using their known pK_a values, the charge at a given pH can be estimated and the pH determined at zero charge. This is done by calculating and summing the partial charges for all acid−base groups by solving a modified Henderson−Hasselbalch eq (eq 1).

Determining a list of pK_a values from protein and peptide sequences containing only natural amino acids is straightforward. However, proteins often undergo post-translational modifications, such as phosphorylations,[3] which will lead to the introduction of novel acidic groups. Additionally, artificial modifications can arise during experimental handling of peptides, such as oxidation[5] and deamidation.[6] Moreover, artificial amino acids and modifications are used during peptide based drug discovery and development.[7,8] If the modifications introduce or remove acidic or basic groups, the pI will be affected to a greater or lesser extent.[6] To be accurate for modified peptides the pI prediction must take all these modifications into account.

Previous algorithms have covered a limited number of modifications such as terminal block,[4,9,10] *N*-acetylneuraminic acid,[4] phosphorylations,[9,11] cyclizations,[12] and methylations.[11] ProMOST has the feature to add extra user defined pK_a values for calculation of custom modifications.[13] The program pIR combines the sequence information with calculation of chemical descriptors and machine learning predictions.[14] However, the algorithms do not have a consistent way of handling the modifications informatically, making them difficult to extend, adapt, and combine.

Here we present an extensible pI prediction algorithm, pICalculaX, which can handle chemical modification effectively by combining bio- and chemoinformatics handling of the molecules. The software has been released as open source on Github (https://github.com/EBjerrum/pICalculax). The code can be used in conjunction with the Python interface to the proprietary Proteax Desktop software[15] for handling sequence to structure conversions. No-cost academic licenses can be requested for Proteax Desktop.

METHODS

pI Prediction Algorithm. The algorithm consists of two major stages. First the molecule is analyzed and the pK_a values[16] of identified acidic or basic groups recorded. Then the pI is predicted by identifying the pH where the sum of

Received: January 16, 2017
Published: July 3, 2017

DOI: 10.1021/acs.jcim.7b00030
J. Chem. Inf. Model. 2017, 57, 1723−1727

partial charges is zero, by solving the Henderson−Hasselbalch equation for the ionization extent (eq 1). Acidic groups need a negative sign of the exponent and nominator.

$$\text{Charge} = \frac{1}{1 + 10^{(\text{pH}-\text{p}Ka_{\text{base1}})}} + \frac{-1}{1 + 10^{-(\text{pH}-\text{p}Ka_{\text{acid1}})}} + ... \quad (1)$$

```
Pseudo code
1. Input molecule or sequence converted to molfile format (Proteax[15])
2. Iterate through rule table
    2.1 If substructure match (RDKit[17])
        2.1.1. Record pKa and charge information from rule
        2.1.2. Substitute matched substructure
3. Use pKa list and Charge list to predict pI.
```

The algorithm works internally with the condensed molfile format.[18−20] Molfile formats describe molecular structures in atomic detail with all atoms and bonds. The Proteax software converts between sequence based formats, full atomic description, and a condensed format in between (illustrated in Figure 1). In the condensed format the natural amino acids are substituted with pseudoatoms, which allows for efficient handling of large molecules (proteins), while still preserving a full description of chemical modifications.

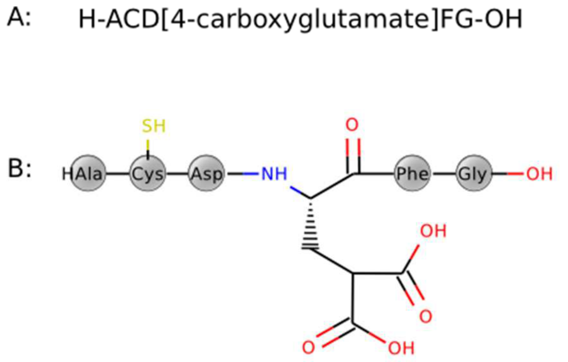


Figure 1. Example of a modified peptide in both protein line notation (PLN) and condensed format. (A) PLN written from the N- to the C-terminus denoted with the hydroxyl of the carboxylic acid (−OH). Letters correspond to natural amino acids, and the modification is noted in square brackets. (B) Drawing of the condensed format. Natural amino acids are pseudoatoms (gray balls), whereas the modified amino acid has explicit molecular structure. Renderings with Proteax and RDKit.

The rule table is built from rules for matching acidic or basic groups, going from specific rules such as amino acid side chains, toward more generic groups such as carboxylic acids. Rules are specified as SMARTS together with their associated pK_a and the charge for the group below the pK_a (0 or 1). Rules allow for more pK_a values and charges to be associated for complex groups. SMARTS are preferably designed to match the atom with the labile hydrogen or lone pair in a specific atomic environment.

RDKit[17] is used for substructure matching between the molecule and the SMARTS. To allow for pseudoatom support, the RDKit source code was patched with an extension to the atomic table file. (patch included in the Supporting Information).

If a rule matches a pseudoatom, the matched pseudoatom is substituted with the glycine pseudoatom, whereas a matched atomic substructure is substituted with atomtype 0.

The Henderson−Hasselbalch equation is used together with the recorded pK_a and charge values to calculate the partial charge of the molecule at a given pH (eq 1). The entire pH range from 0 to 14 is simulated, and the pI is estimated at zero charge. If none, only basic, or only acidic groups are present, the pI cannot be calculated.

Data Sets. The Reaxys database[21,22] was searched for molecules with information about their isoelectric point which also contained a substructure matching the backbone of at least three amino acids. The SD file data set was manually curated and a few molecules removed based on manual judgment of their peptide content. The data set was divided into sets containing peptides with only natural amino acids or modified peptides. Finally, the SD file formatted molecules were converted to protein line notation (PLN) with Proteax and saved into Excel files together with the FASTA sequences. Molecules which failed the conversion to PLN were discarded. The final data sets and the associated PLN modification database are part of the Supporting Information.

Additionally, the data sets from Gauci et al.[9] and the high quality PeptideProphet data set from Heller et al.[23] were downloaded and prepared in FASTA and PLN format accounting for modifications. The average of the pI values assigned by the original authors for each gel fraction was assigned to all peptides from that fraction.

pK_a Sets. Multiple different pK_a value sets have been proposed for usage in peptide pI calculations. Five such sets from Solomons,[24] Lehninger,[25] Grimsley,[16] IPC_peptide,[26] and EMBOSS[27] were tested.

pI Prediction and Comparison. The pI was estimated for all peptides using both the FASTA sequence and the PLN containing the modifications. The correlation coefficient (R^2) and the root-mean-square deviation (RMSD) were calculated between the experimental and the predicted pI. Similar calculations were done with the program pIR,[14,28] with the difference that the PLN was converted to the modification format used by pIR before pI estimation. Plots were made in Python with Matplotlib.

■ RESULTS

For this work, 511 molecules were found in the Reaxys database, which after data preparation resulted in final data sets with 335 unmodified and 99 modified peptides. The algorithm could not calculate the pI of 15 peptides from the data set since these structures contained only basic groups. They were filtered away before further data analysis. The Gauci, Gauci modified, and Heller data sets contained 5757, 250, and 2006 peptides, respectively.

The calculated R^2 and the RMSD for the different data sets using different pK_a sets are summarized in Table 1. The full benchmark matrix with all combinations of pK_a sets, data sets, and methods as well as a comparison with the pIR program can be found in Supplementary Tables 1−3. The best performance is observed with the Grimsley derived pK_a set on the unmodified peptides, with a correlation coefficient of 0.99 and an RMSD is 0.38. The excellent agreement is also evident from Figure 2. However, the test sets with the modified peptides show large differences, depending on whether they are

DOI: 10.1021/acs.jcim.7b00030
J. Chem. Inf. Model. 2017, 57, 1723−1727

Table 1. Calculated Correlation Coefficient (R^2) and Root Mean Square Deviation (RMSD) between Peptide pI Predictions based on FASTA Sequence or Protein Line Notation (PLN) Using Different Datasets and pK_a Sets

	R^2		RMSD	
data set−pK_a set	FASTA	PLN	FASTA	PLN
Reaxys: IPC_peptide[26]				
unmodified peptides	0.97	0.97	0.88	0.88
modified peptides	0.50	0.91	2.84	0.77
Reaxys: EMBOSS[27,29]				
unmodified peptides	0.99	0.99	0.66	0.66
modified peptides	0.45	0.92	2.92	0.76
Reaxys: Solomons[24]				
unmodified peptides	0.96	0.96	0.89	0.89
modified peptides	0.50	0.91	2.83	0.77
Reaxys: Lehninger[25]				
unmodified peptides	0.96	0.96	0.91	0.91
modified peptides	0.51	0.90	2.82	0.81
Reaxys: Grimsley[16]				
unmodified peptides	0.99	0.99	0.38	0.38
modified peptides	0.45	0.91	3.30	0.90
Gauchi: Grimsley				
unmodified peptides	0.90	0.91	0.32	0.33
modified peptides	0.53	0.85	1.83	0.55
Heller: Grimsley				
unmodified peptides	0.90	0.94	0.61	0.49

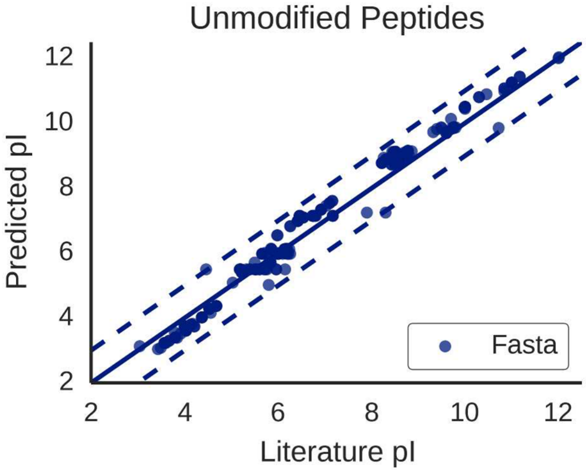


Figure 2. Scatter plot of Reaxys literature pI values and values predicted from FASTA sequence alone for peptides containing only natural amino acids. The Grimsley[16] pK_a set was used. The solid blue line marks the perfect correlation with the dashed lines at ±1 pH unit.

predicted from PLN or FASTA sequence. The results for using the FASTA sequences show no good correlation between data set pI and predicted pI for the Grimsley pK_a set, and many outliers can be seen in Figure 3, Graph A (R^2 of 0.45 and RMSD of 3.30).

By taking the modifications into account with the PLN notation, the excellent agreements between theoretical and literature values are restored to some degree (R^2 of 0.91 and RMSD of 0.90 for the Grimsley pK_a set and Figure 3, Graph B). The best performance on the modified peptides was seen with the EMBOSS pK_a set[29] using the PLN data set, with an R^2 of 0.92 and RMSD of 0.76, which is on par with the performance of the IPC_peptide pK_a set.[26]

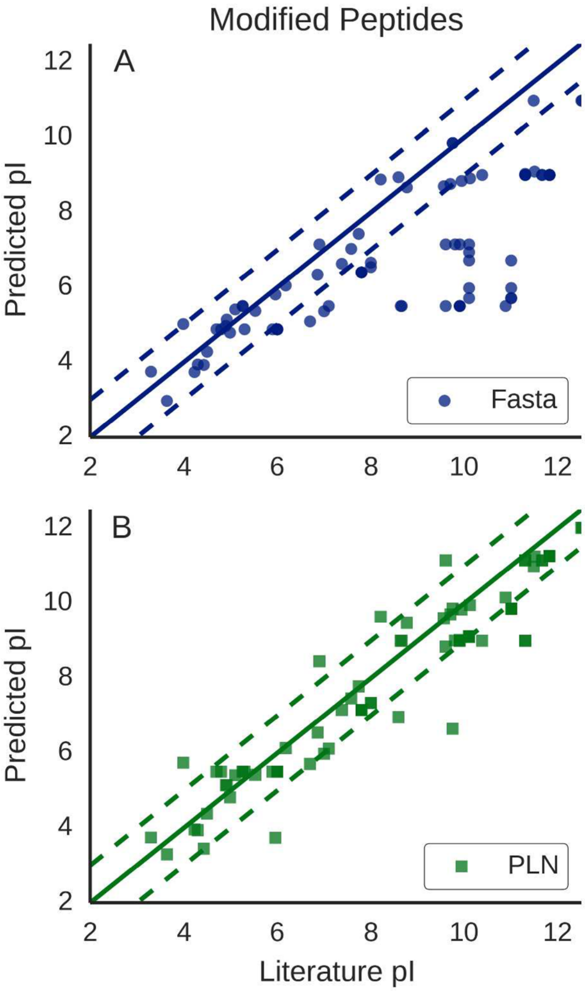


Figure 3. Scatter plots between Reaxys literature pI values and predicted pI values for peptides with modifications or unnatural amino acids. The Grimsley[16] pK_a set was used. (A) Predictions made from FASTA sequence (blue). (B) Predictions made from protein line notaion (PLN) which takes modifications into account (green). The solid lines mark the perfect correlation with the dashed lines at ±1 pH unit.

DISCUSSION

The results show the importance of taking chemical modifications into account when predicting pI values. However, the predicted pI was changed less than 1 pH unit for 49 of the 99 peptides in the Reaxys data set with modifications. Modifications that do not change or introduce an acidic or basic group will not affect the calculations, and even if a group is changed, it can have little effect if the pK_a value is far away from the pI value and the charge after the modification remains the same.

During analysis, a group of peptides failed to produce pI values during prediction from the PLN sequence. They contained only basic groups, but the original reference showed that they had been assigned a pI value of 14 without any description of the experimental procedure.[30] The values could be based on pI prediction or the value should have been assigned as >14. For this analysis, the effect of predicted pI values is not detrimental as the focus is on the consequences of not including the modifications in the calculations. However, predicted pI values must not be used for tuning of pK_a value sets and pI prediction algorithms.[26]

DOI: 10.1021/acs.jcim.7b00030
J. Chem. Inf. Model. 2017, 57, 1723−1727

The Gauci[9] and Heller[23] data sets are derived from shotgun proteomics data. Isoelectric focusing were used to fractionate the digested protein samples before further separation and identification using MS-MS. During experimental procedures the free cysteine side chains are blocked with iodoacetamide[9,23] which could lead to problems for calculations on peptides containing free cysteines. Moreover, during data filtering the predicted value of the peptide sequence is matched to the gel fraction, and hits outside two standard deviations are removed. This biases the data set with peptides that can be predicted with the pI prediction algorithm used in the proteomics study. These kinds of experimental, but filtered, data sets must not be used for further tuning of pK_a value sets and pI prediction algorithms[26] as the tuned pK_a values will regress toward the pK_a values used to filter the data sets.

pICalculax was tested with five different pK_a sets. The pK_a values from Grimsley[16] are derived from protein NMR titration experiments, whereas the IPC_peptide pK_a set has been tuned to give good pI predictions against experimental data sets.[26] The best performance for the unmodified peptides was observed with the Grimsley pK_a set, which had otherwise showed below average performance in a recent benchmark,[26] whereas a previous benchmark[28] had shown good performance. A possible reason for the bad performance in the most recent benchmark[26] is that it was done on a data set derived from proteomics data; these are problematic to use for peptide pI algorithm tuning and benchmarking (see above). Moreover, the benchmark seems to have been done[26] on the predicted values from the Heller SEQUEST and PHENYX data sets,[23] rather than on the average pH value of the IEF gel fractions. The results of the benchmark on the Heller and Gauci data sets should thus be interpreted cautiously and can fully explain the differences in performance observed between the studies.

The pIR[28] program performed equally well as pICalculax on the Gauci modified, but not the Reaxys modified data set (Supplementary Table 3). This may be because the Gauci modified data set only contains phosphorylations and *n*-terminal acetylations which are modifications known by pIR. The Reaxys data set contains multiple other modifications which could not readily be converted to a format understood by pIR and thus not accounted for in the pI calculation.

The performance on the Reaxys modified data set predicted from PLN was better using the pK_a sets IPC_peptide and EMBOSS. This suggests that a performance gain could be derived from a pK_a set tuning, balancing the performance on modified and unmodified peptides if a suitable data set could collected. The default pK_a set used in pICalculax is from Grimsley, but the other pK_a sets are included and can be used instead through a setting in the program source code.

The current implementation of the algorithm uses the naïve approach that the extent of protonation for each group is independent of each other, and this could lead to wrong predictions for peptides with many neighboring acidic and basic groups or with acidic or basic groups at the termini.

CONCLUSION

We have presented pICalculax, a program which can be used to include modifications of peptides in pI calculations. The approach is extensible to new modifications or artificial amino acids, as the rule table can be updated with new SMARTS rules and pK_a values as the need dictates. The use of PLN notation simplifies data management, as there is no need for storing modifications and sequence information separately.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.7b00030.

Supplementary tables (PDF)

Information about the final Reaxys data sets and the predicted pI values and the modification database for Proteax Desktop and a patch file for RDKit to support condensed molfile formats of peptides (ZIP)

AUTHOR INFORMATION

Corresponding Author

*E-mail: esben@wildcardconsulting.dk.

ORCID

Esben J. Bjerrum: 0000-0003-1614-7376

Funding

The research and publication were completely funded by the authors' respective companies.

Notes

The authors declare the following competing financial interest(s): J.H.J. is the owner of Biochemfusion ApS which licenses Proteax Desktop used together with pICalculax in this study. The authors E.J.B. and J.H.J. are employed by Wildcard Pharmaceutical Consulting and Biochemfusion ApS, respectively. Both companies are usually contracted by biotechnology/pharma companies to provide third party services.

ABBREVIATIONS

pH, potential of hydrogen; pI, isoelectric point; pK_a, acid dissociation constant; PLN, protein line notation; RMSD, root mean square deviations; SD file, structure-data file; SMARTS, Smiles arbitrary target specification

REFERENCES

(1) Righetti, P. G.; Fasoli, E.; Righetti, S. C. Conventional Isoelectric Focusing. In Gel Slabs and Capillaries and Immobilized pH Gradients. *Methods Biochem. Anal.* **2011**, *54*, 379−409.

(2) Righetti, P. G.; Sebastiano, R.; Citterio, A. Capillary Electrophoresis and Isoelectric Focusing in Peptide and Protein Analysis. *Proteomics* **2013**, *13*, 325−340.

(3) Schuurmans Stekhoven, F. M. A. H.; Gorissen, M. H. A. G.; Flik, G. The Isoelectric Point, a Key to Understanding a Variety of Biochemical Problems: a Minireview. *Fish Physiol. Biochem.* **2008**, *34*, 1−8.

(4) Bjellqvist, B.; Hughes, G. J.; Pasquali, C.; Paquet, N.; Ravier, F.; Sanchez, J. C.; Frutiger, S.; Hochstrasser, D. The Focusing Positions of Polypeptides in Immobilized pH Gradients Can be Predicted from Their Amino Acid Sequences. *Electrophoresis* **1993**, *14*, 1023−1031.

(5) Perdivara, I.; Deterding, L. J.; Przybylski, M.; Tomer, K. B. Mass Spectrometric Identification of Oxidative Modifications of Tryptophan Residues in Proteins: Chemical Artifact or Post-Translational Modification? *J. Am. Soc. Mass Spectrom.* **2010**, *21*, 1114−1117.

(6) Lengqvist, J.; Eriksson, H.; Gry, M.; Uhlén, K.; Björklund, C.; Bjellqvist, B.; Jakobsson, P.-J.; Lehtiö, J. Observed Peptide pI and Retention Time Shifts as a Result of Post-Translational Modifications in Multidimensional Separations Using Narrow-Range IPG-IEF. *Amino Acids* **2011**, *40*, 697−711.

(7) Gentilucci, L.; De Marco, R.; Cerisoli, L. Chemical Modifications Designed to Improve Peptide Stability: Incorporation of Non-Natural Amino Acids, Pseudo-Peptide Bonds, and Cyclization. *Curr. Pharm. Des.* **2010**, *16*, 3185−3203.

(8) Fosgerau, K.; Hoffmann, T. Peptide Therapeutics: Current Status and Future Directions. *Drug Discovery Today* **2015**, *20*, 122−128.

(9) Gauci, S.; van Breukelen, B.; Lemeer, S. M.; Krijgsveld, J.; Heck, A. J. R. A Versatile Peptide pI Calculator for Phosphorylated and N-Terminal Acetylated Peptides Experimentally Tested using Peptide Isoelectric Focusing. *Proteomics* **2008**, *8*, 4898−4906.

(10) Lear, S.; Cobb, S. L. Pep-Calc.com: a Set of Web Utilities for the Calculation of Peptide and Peptoid Properties and Automatic Mass Spectral Peak Assignment. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 271−277.

(11) Hung, C.-W.; Kübler, D.; Lehmann, W. D. pI-Based Phosphopeptide Enrichment Combined with nanoESI-MS. *Electrophoresis* **2007**, *28*, 2044−2052.

(12) Zhang, X.; Højrup, P. Cyclization of the N-terminal X-Asn-Gly Motif During Sample Preparation for Bottom-Up Proteomics. *Anal. Chem.* **2010**, *82*, 8680−8685.

(13) Halligan, B. D.; Ruotti, V.; Jin, W.; Laffoon, S.; Twigger, S. N.; Dratz, E. A. ProMoST (Protein Modification Screening Tool): a Web-Based Tool for Mapping Protein Modifications on Two-Dimensional Gels. *Nucleic Acids Res.* **2004**, *32*, W638−W644.

(14) Perez-Riverol, Y.; Audain, E.; Millan, A.; Ramos, Y.; Sanchez, A.; Vizcaíno, J. A.; Wang, R.; Müller, M.; Machado, Y. J.; Betancourt, L. H.; González, L. J.; Padrón, G.; Besada, V. Isoelectric Point Optimization Using Peptide Descriptors and Support Vector Machines. *J. Proteomics* **2012**, *75*, 2269−2274.

(15) Jensen, J. H. Proteax Desktop. http://www.biochemfusion.com/downloads/#ProteaxDesktop (accessed Dec 22, 2016).

(16) Grimsley, G. R.; Scholtz, J. M.; Pace, C. N. A Summary of the Measured pK Values of the Ionizable Groups in Folded Proteins. *Protein Sci.* **2008**, *18*, 247−251.

(17) Landrum, G. A. RDKit: Open-Source Cheminformatics Software. http://www.rdkit.org/, https://github.com/rdkit/rdkit (accessed Dec 22, 2016).

(18) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. *J. Chem. Inf. Model.* **1992**, *32*, 244−255.

(19) Jensen, J. H.; Hoeg-Jensen, T.; Padkjaer, S. B. Building a BioChemformatics Database. *J. Chem. Inf. Model.* **2008**, *48*, 2404−2413.

(20) CTfile Formats. http://accelrys.com/products/informatics/cheminformatics/ctfile-formats/no-fee.php (accessed Dec 22, 2016).

(21) Goodman, J. Computer Software Review: Reaxys. *J. Chem. Inf. Model.* **2009**, *49*, 2897−2898.

(22) Reaxys. http://www.reaxys.com (accessed Oct 15, 2016).

(23) Heller, M.; Ye, M.; Michel, P. E.; Morier, P.; Stalder, D.; Jünger, M. A.; Aebersold, R.; Reymond, F.; Rossier, J. S. Added Value for Tandem Mass Spectrometry Shotgun Proteomics Data Validation Through Isoelectric Focusing of Peptides. *J. Proteome Res.* **2005**, *4*, 2273−2282.

(24) Solomons, T. W. G.; Fryhle, C. B.; Snyder, S. A. *Organic Chemistry*; John Wiley & Sons, USA, 1992.

(25) Nelson, D.; Lehninger, A. L.; Cox, M. *Lehninger Principles of Biochemistry*; Macmillan learning, USA: New York, 2008.

(26) Kozlowski, L. P. IPC - Isoelectric Point Calculator. *Biol. Direct* **2016**, *11*, 55.

(27) Rice, P.; Longden, I.; Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **2000**, *16*, 276−277.

(28) Audain, E.; Ramos, Y.; Hermjakob, H.; Flower, D. R.; Perez-Riverol, Y. Accurate Estimation of Isoelectric Point of Protein and Peptide Based on Amino Acid Sequences. *Bioinformatics* **2016**, *32*, 821−827.

(29) EMBOSS Iep. http://emboss.sourceforge.net/apps/release/6.6/emboss/apps/iep.html (accessed Mar 09, 2017).

(30) Cui, Y.; Pattabiraman, A.; Lisko, B.; Collins, S. C.; McAlpine, M. C. Recognition of Patterned Molecular Ink with Phage Displayed Peptides. *J. Am. Chem. Soc.* **2010**, *132*, 1204−1205.