PhoglyStruct: Prediction of phosphoglycerylated lysine residues using structural properties of amino acids

Abel Chandra (abelavit@gmail.com)
University of the South Pacific
Alok Sharma (alok.sharma@griffith.edu.au)
RIKEN Center for Integrative Medical Sciences

Method Article

Keywords: Post-translational modification, Phosphoglycerylation, Structural Properties, Predictor

Posted Date: August 22nd, 2019

DOI: https://doi.org/10.21203/rs.2.1673/v1

Version of Record: A version of this preprint was published at Scientific Reports on December 1st, 2018.
See the published version at https://doi.org/10.1038/s41598-018-36203-8.

Abstract

The biological process known as post-translational modification \(PTM) contributes to diversifying the proteome hence affecting many aspects of normal cell biology and pathogenesis. There have been many recently reported PTMs, but lysine phosphoglycerylation has emerged as the most recent subject of interest. Despite a large number of proteins being sequenced, the experimental method for detection of phosphoglycerylated residues remains an expensive, time-consuming and inefficient endeavor in the post-genomic era. Instead, the computational methods are being proposed for accurately predicting phosphoglycerylated lysines. Though a number of predictors are available, performance in detecting phosphoglycerylated lysine residues is still limited. In this paper, we propose a new predictor called PhoglyStruct that utilizes structural information of amino acids alongside a multilayer perceptron classifier for predicting phosphoglycerylated and non-phosphoglycerylated lysine residues. For the experiment, we located phosphoglycerylated and non-phosphoglycerylated lysines in our employed benchmark. We then derived and integrated properties such as accessible surface area, backbone torsion angles, and local structure conformations. PhoglyStruct showed significant improvement in the ability to detect phosphoglycerylated residues from non-phosphoglycerylated ones when compared to previous predictors. The sensitivity, specificity, accuracy, Mathews correlation coefficient and AUC were 0.8542, 0.7597, 0.7834, 0.5468 and 0.8077, respectively. The data and Matlab/Octave software packages are available at https://github.com/abelavit/PhoglyStruct.

Procedure

The codes in this repository are in two categories. One is based on commercial software \(Matlab), while the other on non-commercial software \(Octave). - The train and test datasets used for implementing PhoglyStruct are .mat files by the names 'train' and 'test' respectively \(the three features namely tau, pc and ph are not present in train and test but in the .mat files 'original_train' and 'original_test' has all the features which can be viewed for reference). - In these datasets, the first column is the protein sequence name, second column the feature vector, label in the third \('1' for phosphoglycerylated and '0' for non-phosphoglycerylated), and the fourth column indicates the amino acid number where lysine K is located in the protein sequence. - The dataset with removed features \(tau, pc and ph) were converted to arff files using the .m file named 'removed_features_arff'. The datasets were converted to arff files to train multilayer perceptron on WEKA \(arff files can be found in 'PhoglyStruct_arffs' folder). - The algorithm \(.m file) used for generating the original train and original test datasets is called 'PhoglyStruct'. These datasets were also generated for the CKSAAP_PhoglySite method containing CKSAAP features and its arff files were used to train multilayer perceptron on WEKA for comparison \(arff files can be found in 'CKSAAP_arffs' folder). - The algorithm \(.m file) used for generating the test and train datasets is called 'CKSAAP'. The performance of test set was also obtained for Phogly-PseAAC and iPGK-PseAAC method by comparing the lysine k predictions when FASTA format of the protein sequence was uploaded to its webservers. - The Phogly-PseAAC predictions of all lysine k is stored in .mat file named 'Phogly_PseAAC_Result' while for iPGK-PseAAC in 'iPGK_PseAAC_result'. Since these two methods were

not implemented in our work, arff files for these methods is not generated so the result is obtained by executing the .m file named 'Phogly_PseAAC' for Phogly-PseAAC method and .m file named 'iPGK_PseAAC' for iPGK-PseAAC method that calculate performance based on predictions carried out on the respective webservers. - Moreover, Phoglystruct features were compared to a simpler set of features that assigns a value of 1 when the amino acid at the particular position in the peptide P matches with one of the amino acids of the genome while a value of 0 is assigned to the rest of the amino acids. The resulting matrix obtained is of a 20x5 dimension. This matrix is converted into a 100-dimensional feature vector representing each lysine residue. This simple feature for the corresponding PhoglyStruct's train and test datasets was constructed by executing the code \(.m file) named 'Simple_Feature'. The arff files generated were used to train MLP on WEKA. Please see details on training MLP on WEKA and obtaining AUC below. WEKA and AUC calculation details: - 10-fold cross-validation of our method and CKSAAP_PhoglySite method is carried out using the arff files. The WEKA version 3.8.2 was used in this work. On WEKA, open file to train and on the classifier tab, choose MultilayerPerceptron under functions. The parameters of MultilayerPerceptron are kept as default. Supply the corresponding test set. Please also choose csv format for output predictions under 'more options'. - After training is complete, use the confusion matrix to calculate the performance metrics sensitivity, specificity, G-Mean, accuracy, mcc and F-Measure \(the excel file named 'MLP WEKA metric calculator' can be used for calculation). For calculating AUC, copy and paste the predictions on test set into a txt file \(the predictions on test set are provided for PhoglyStruct, CKSAAP_PhoglySite, iPGK-PseAAC and Phogly-PseAAC methods by the names 'AUC_Data_PhoglyStruct', 'AUC_Data_CKSAAP', 'AUC_Data_iPGK_PseAAC' and 'AUC_Data_Phogly_PseAAC' respectively). - The data for calculating AUC of the method that utilizes simple features is also provided by the name 'AUC_Data_Simple_Features'. To calculate the AUC, please execute the .R file named 'Calculating_AUC'. Footnotes: To find in detail the CKSAAP_PhoglySite feature extraction method for each lysine k, please see the .m file named ‘CKSAAP_Preprocessing’. After the code execution, features are saved in the Final_Data variable. Final_Data is the same file used when comparing for the CKSAAP_PhoglySite method. To verify the algorithm for calculating the CKSAAP features, code named 'CKSAAP_Preprocessing_Xu_Dataset' was developed to run on Xu's Dataset and the feature rank achieved by this alogorithm was compared to the rank achieved in CKSAAP_PhoglySite work and they come to the same ranking. The rank achieved by CKSAAP_PhoglySite method is highlighted in table 3 of their paper. The file also contains FASTA format of the phosphoglycerylation dataset which was used to obtained the predictions of all lysine k from the Phogly–PseAAC webserver accessible at http://app.aporc.org/Phogly-PseAAC/ and iPGK-PseAAC webserver accessible at http://app.aporc.org/iPGK-PseAAC/

Anticipated Results

- The result shows a significant improvement in the ability to detect phosphoglycerylated residues from non-phosphoglycerylated ones when compared to previous predictors. - The sensitivity, specificity, accuracy, Mathews correlation coefficient and AUC were 0.8542, 0.7597, 0.7834, 0.5468 and 0.8077, respectively.

References

1. Huang, J., et al., Enrichment and separation techniques for large-scale proteomics analysis of the protein post-translational modifications. Journal of Chromatography A, 2014. 1372: p. 1-17. 2. Lanouette, S., et al., The functional diversity of protein lysine methylation. Molecular systems biology, 2014. 10\(4): p. 724. 3. Liu, Z., et al., CPLM: a database of protein lysine modifications. Nucleic acids research, 2014. 42\(D1): p. D531-D536. 4. Choudhary, C., et al., Lysine acetylation targets protein complexes and co-regulates major cellular functions. Science, 2009. 325\(5942): p. 834-840. 5. Johansen, M.B., L. Kiemer, and S. Brunak, Analysis and prediction of mammalian protein glycation. Glycobiology, 2006. 16\(9): p. 844-853. 6. Park, J., et al., SIRT5-mediated lysine desuccinylation impacts diverse metabolic pathways. Molecular cell, 2013. 50\(6): p. 919-930. 7. Tan, M., et al., Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. Cell, 2011. 146\(6): p. 1016-1028. 8. Lan, F. and Y. Shi, Epigenetic regulation: methylation of histone and non-histone proteins. Science in China Series C: Life Sciences, 2009. 52\(4): p. 311-322. 9. Cheng, Z., et al., Molecular characterization of propionyllysines in non-histone proteins. Molecular & Cellular Proteomics, 2009. 8\(1): p. 45-52. 10. Iyer, L.M., A.M. Burroughs, and L. Aravind, Unraveling the biochemistry and provenance of pupylation: a prokaryotic analog of ubiquitination. Biology direct, 2008. 3\(1): p. 45. 11. Szondy, Z., et al., Transglutaminase 2 in human diseases. BioMedicine, 2017. 7\(3). 12. Li, S., et al., Loss of post-translational modification sites in disease, in Biocomputing 2010. 2010, World Scientific. p. 337-347. 13. Liddy, K.A., M.Y. White, and S.J. Cordwell, Functional decorations: post-translational modifications and heart disease delineated by targeted proteomics. Genome medicine, 2013. 5\(2): p. 20. 14. Spinelli, F.R., et al., Post-translational modifications in rheumatoid arthritis and atherosclerosis: Focus on citrullination and carbamylation. Journal of International Medical Research, 2016. 44\(1_suppl): p. 81-84. 15. Ju, Z., J.-Z. Cao, and H. Gu, Predicting lysine phosphoglycerylation with fuzzy SVM by incorporating k-spaced amino acid pairs into Chou s general PseAAC. Journal of Theoretical Biology, 2016. 397: p. 145-150. 16. Moellering, R.E. and B.F. Cravatt, Functional lysine modification by an intrinsically reactive primary glycolytic metabolite. Science, 2013. 341\(6145): p. 549-553. 17. Bulcun, E., M. Ekici, and A. Ekici, Disorders of glucose metabolism and insulin resistance in patients with obstructive sleep apnoea syndrome. International journal of clinical practice, 2012. 66\(1): p. 91-97. 18. Kolwicz Jr, S.C. and R. Tian, Glucose metabolism and cardiac hypertrophy. Cardiovascular research, 2011. 90\(2): p. 194-201. 19. Dehzangi, A., et al., PSSM-Suc: Accurately predicting succinylation using position specific scoring matrix into bigram for feature extraction. Journal of Theoretical Biology, 2017. 425: p. 97-102. 20. Chou, K.-C. and H.-B. Shen, Recent progress in protein subcellular location prediction. Analytical Biochemistry, 2007. 370\(1): p. 1-16. 21. Jia, J., et al., iSuc-PseOpt: identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset. Analytical Biochemistry, 2016. 497: p. 48-56. 22. López, Y., et al., Success: evolutionary and structural properties of amino acids prove effective for succinylation site prediction. BMC Genomics, 2018. 19\(1): p. 923. 23. Ju, Z. and J.-J. He, Prediction of lysine propionylation sites using biased SVM and incorporating four different sequence features into Chou's PseAAC. Journal of Molecular Graphics and Modelling, 2017. 76: p. 356-363. 24. Xu, Y., et al., Mal-Lys: prediction of lysine malonylation sites in

proteins integrated sequence-based features with mRMR feature selection. Scientific reports, 2016. 6: p. 38318. 25. Xiang, Q., et al., Prediction of Lysine Malonylation Sites Based on Pseudo Amino Acid. Combinatorial chemistry & high throughput screening, 2017. 20\(7): p. 622-628. 26. Du, Y., et al., Prediction of Protein Lysine Acylation by Integrating Primary Sequence Information with Multiple Functional Features. Journal of proteome research, 2016. 15\(12): p. 4234-4244. 27. Qiu, W.-R., et al., iUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a gray system model. Journal of Biomolecular Structure and Dynamics, 2015. 33\(8): p. 1731-1742. 28. Hou, T., et al., LAceP: lysine acetylation site prediction using logistic regression classifiers. PloS one, 2014. 9\(2): p. e89575. 29. Jia, J., et al., pSumo-CD: predicting sumoylation sites in proteins with covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC. Bioinformatics, 2016. 32\(20): p. 3133-3141. 30. Qiu, W.-R., et al., iKcr-PseEns: Identify lysine crotonylation sites in histone proteins with pseudo components and ensemble classifier. Genomics, 2017. 31. Ju, Z. and H. Gu, Predicting pupylation sites in prokaryotic proteins using semi-supervised self-training support vector machine algorithm. Analytical biochemistry, 2016. 507: p. 1-6. 32. Bakhtiarizadeh, M.R., et al., Neural network and SVM classifiers accurately predict lipid binding proteins, irrespective of sequence homology. Journal of Theoretical Biology, 2014. 356: p. 213-222. 33. Liu, Y., et al., PTM-ssMP: A Web Server for Predicting Different Types of Post-translational Modification Sites Using Novel Site-specific Modification Profile. International Journal of Biological Sciences, 2018. 14\(8): p. 946-956. 34. Wang, B., M. Wang, and A. Li, Prediction of post-translational modification sites using multiple kernel support vector machine. PeerJ, 2017. 5: p. e3261. 35. Fan, W., et al., Prediction of protein kinase-specific phosphorylation sites in hierarchical structure using functional information and random forest. Amino acids, 2014. 46\(4): p. 1069-1078. 36. Xu, Y., et al., Phogly–PseAAC: prediction of lysine phosphoglycerylation in proteins incorporating with position-specific propensity. Journal of Theoretical Biology, 2015. 379: p. 10-15. 37. Chen, Q.-Y., J. Tang, and P.-F. Du, Predicting protein lysine phosphoglycerylation sites by hybridizing many sequence based features. Molecular BioSystems, 2017. 13\(5): p. 874-882. 38. Liu, L.-M., Y. Xu, and K.-C. Chou, iPGK-PseAAC: identify lysine phosphoglycerylation sites in proteins by incorporating four different tiers of amino acid pairwise coupling information into the general PseAAC. Medicinal Chemistry, 2017. 13\(6): p. 552-559. 39. Sharma, A., et al., A strategy to select suitable physicochemical attributes of amino acids for protein fold recognition. BMC bioinformatics, 2013. 14\(1): p. 233. 40. Li, W. and A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics, 2006. 22\(13): p. 1658-1659. 41. Dehzangi, A., et al., Improving succinylation prediction accuracy by incorporating the secondary structure via helix, strand and coil, and evolutionary information from profile bigrams. PloS one, 2018. 13\(2): p. e0191900. 42. Liu, Z., et al., iDNA-Methyl: Identifying DNA methylation sites via pseudo trinucleotide composition. Analytical biochemistry, 2015. 474: p. 69-77. 43. Jia, J., et al., iPPBS-Opt: a sequence-based ensemble classifier for identifying protein-protein binding sites by optimizing imbalanced training datasets. Molecules, 2016. 21\(1): p. 95. 44. Heffernan, R., et al., Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. Scientific reports, 2015. 5: p. 11476. 45. Lyons, J., et al., Predicting backbone Cα angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network.

Journal of computational chemistry, 2014. 35\(28): p. 2040-2046. 46. Heffernan, R., et al., Highly accurate sequence-based prediction of half-sphere exposures of amino acid residues in proteins. Bioinformatics, 2015. 32\(6): p. 843-849. 47. Yang, Y., et al., SPIDER2: A Package to Predict Secondary Structure, Accessible Surface Area, and Main-Chain Torsional Angles by Deep Neural Networks. Prediction of Protein Secondary Structure, 2017: p. 55-63. 48. Faraggi, E., et al., SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. Journal of computational chemistry, 2012. 33\(3): p. 259-267. 49. McGuffin, L.J., K. Bryson, and D.T. Jones, The PSIPRED protein structure prediction server. Bioinformatics, 2000. 16\(4): p. 404-405. 50. Faraggi, E., et al., Predicting continuous local structure and the effect of its substitution for secondary structure in fragment-free protein structure prediction. Structure, 2009. 17\(11): p. 1515-1527. 51. Taherzadeh, G., et al., Sequence-based prediction of protein–carbohydrate binding sites using support vector machines. Journal of chemical information and modeling, 2016. 56\(10): p. 2115-2122. 52. Taherzadeh, G., et al., Sequence-based prediction of protein–peptide binding sites using support vector machine. Journal of computational chemistry, 2016. 37\(13): p. 1223-1229. 53. López, Y., et al., SucStruct: Prediction of succinylated lysine residues by using structural properties of amino acids. Analytical Biochemistry, 2017. 527: p. 24-32. 54. Dehzangi, A., et al. Enhancing protein fold prediction accuracy using evolutionary and structural features. in IAPR International Conference on Pattern Recognition in Bioinformatics. 2013. Springer. 55. Sharma, R., et al., OPAL: Prediction of MoRF regions in intrinsically disordered protein sequences. Bioinformatics, 2018. 56. Uddin, M.R., et al., EvoStruct-Sub: An accurate Gram-positive protein subcellular localization predictor using evolutionary and structural features. Journal of theoretical biology, 2018. 443: p. 138-146. 57. Lins, L., A. Thomas, and R. Brasseur, Analysis of accessible surface of residues in proteins. Protein science, 2003. 12\(7): p. 1406-1417. 58. Pan, B.-B., et al., 3D structure determination of a protein in living cells using paramagnetic NMR spectroscopy. Chemical Communications, 2016. 52\(67): p. 10237-10240. 59. Dor, O. and Y. Zhou, Real-SPINE: An integrated system of neural networks for real-value prediction of protein structural properties. PROTEINS: Structure, Function, and Bioinformatics, 2007. 68\(1): p. 76-81. 60. Xue, B., et al., Real-value prediction of backbone torsion angles. Proteins: Structure, Function, and Bioinformatics, 2008. 72\(1): p. 427-433. 61. Hall, M., et al., The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 2009. 11\(1): p. 10-18. 62. Hamada, M. and K. Asai, A classification of bioinformatics algorithms from the viewpoint of maximizing expected accuracy \(MEA). Journal of Computational Biology, 2012. 19\(5): p. 532-549. 63. Matthews, B.W., Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta \(BBA)-Protein Structure, 1975. 405\(2): p. 442-451. 64. Chou, K.-C. and C.-T. Zhang, Prediction of protein structural classes. Critical reviews in biochemistry and molecular biology, 1995. 30\(4): p. 275-349. 65. Chou, K.C., Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins: Structure, Function, and Bioinformatics, 2001. 43\(3): p. 246-255. 66. Chou, K.-C., Some remarks on protein attribute prediction and pseudo amino acid composition. Journal of theoretical biology, 2011. 273\(1): p. 236-247. 67. Kabir, M. and M. Hayat, iRSpot-GAEnsC: identifing recombination spots via ensemble classifier and extending the concept of Chou's PseAAC to formulate DNA samples. Molecular genetics and genomics, 2016. 291\(1): p. 285-296. 68. Khan, M., et al., Unb-DPC: Identify mycobacterial membrane protein types by

incorporating un-biased dipeptide composition into Chou's general PseAAC. Journal of theoretical biology, 2017. 415: p. 13-19. 69. Meher, P.K., et al., Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. Scientific reports, 2017. 7: p. 42362. 70. Tripathi, P. and P.N. Pandey, A novel alignment-free method to classify protein folding types by combining spectral graph clustering with Chou's pseudo amino acid composition. Journal of theoretical biology, 2017. 424: p. 49-54. 71. Xiao, X., et al., iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking via the benchmark dataset optimization approach. J Biomol Struct Dyn \(JBSD), 2015. 33\(10): p. 2221-2233. 72. Shatabda, S., et al., iPHLoc-ES: Identification of bacteriophage protein locations using evolutionary and structural features. Journal of theoretical biology, 2017. 435: p. 229-237. 73. Dehzangi, A. and S. Karamizadeh, Solving protein fold prediction problem using fusion of heterogeneous classifiers. International Information Institute \(Tokyo). Information, 2011. 14\(11): p. 3611. 74. Zhang, N., et al., Discriminating between lysine sumoylation and lysine acetylation using mRMR feature selection and analysis. PloS one, 2014. 9\(9): p. e107464. 75. Li, B.-Q., et al., Prediction of protein cleavage site with feature selection by random forest. PloS one, 2012. 7\(9): p. e45854. 76. Li, B.-Q., et al., Prediction of protein domain with mRMR feature selection and analysis. PLoS One, 2012. 7\(6): p. e39308. 77. Brandes, N., D. Ofer, and M. Linial, ASAP: a machine learning framework for local protein properties. Database, 2016. 2016. 78. Song, J., et al., PhosphoPredict: A bioinformatics tool for prediction of human kinase-specific phosphorylation substrates and sites by integrating heterogeneous feature selection. Scientific Reports, 2017. 7\(1): p. 6862. 79. Chou, K.-C. and H.-B. Shen, Recent advances in developing web-servers for predicting protein attributes. Natural Science, 2009. 1\(02): p. 63. 80. Chen, W., et al., iRNA-AI: identifying the adenosine to inosine editing sites in RNA sequences. Oncotarget, 2017. 8\(3): p. 4208. 81. Cheng, X., X. Xiao, and K.-C. Chou, pLoc-mGneg: Predict subcellular localization of Gram-negative bacterial proteins by deep gene ontology learning via general PseAAC. Genomics, 2017. 82. Cheng, X., et al., pLoc-mAnimal: predict subcellular localization of animal proteins with both single and multiple sites. Bioinformatics, 2017. 33\(22): p. 3524-3531. 83. Liu, B., et al., iRSpot-EL: identify recombination spots with an ensemble learning approach. Bioinformatics, 2016. 33\(1): p. 35-41. 84. Liu, B., F. Yang, and K.-C. Chou, 2L-piRNA: a two-layer ensemble classifier for identifying piwi-interacting RNAs and their function. Molecular Therapy-Nucleic Acids, 2017. 7: p. 267-277. 85. Cheng, X., X. Xiao, and K.-C. Chou, pLoc-mEuk: Predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general PseAAC. Genomics, 2017. 86. Ehsan, A., et al., A Novel Modeling in Mathematical Biology for Classification of Signal Peptides. Scientific reports, 2018. 8\(1): p. 1039. 87. Feng, P., et al., iDNA6mA-PseKNC: Identifying DNA N6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. Genomics, 2018. 88. Liu, B., et al., iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. Bioinformatics, 2017. 34\(1): p. 33-40. 89. Song, J., et al., PREvaIL, an integrative approach for inferring catalytic residues using sequence, structural, and network features in a machine-learning framework. Journal of theoretical biology, 2018. 443: p. 125-137. 90. Chou, K.C., An Unprecedented Revolution in Medicinal Chemistry Driven by the Progress of Biological Science. Current Topics in Medicinal Chemistry, 2017. 17\(21): p.

2337-2358. 91. Chou, K.-C., Impacts of bioinformatics to medicinal chemistry. Medicinal chemistry, 2015. 11\(3): p. 218-234.

Acknowledgements

This research was partially supported by JST CREST Grant Number JPMJCR1412, Japan, and JSPS KAKENHI Grant Numbers 17H06307 and 17H06299, Japan, and Nanken-Kyoten, TMDU, Japan. We would also like to acknowledge the reviewers of Scientific Reports for their constructive comments.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplement0.rar