

基于集成模型的蛋白变构位点预测方法

乔仕杰, 胡芳睿, 李春华*

北京工业大学化学与生命科学学院, 北京

收稿日期: 2024年8月22日; 录用日期: 2024年9月15日; 发布日期: 2024年9月24日

摘要

变构是调节蛋白质功能的重要机制, 对许多生物过程至关重要。变构调节剂比正构剂具有更高的特异性和更低的毒副作用, 这使得变构药物设计比正构药物设计有更多的优势。变构位点的发现是变构药物设计的前提, 目前实验上获得的变构位点多是偶然所得, 因此亟待发展有效的理论方法来预测蛋白质变构位点。本工作提出了一种集成的机器学习方法AllosEC用于预测蛋白质变构口袋, 该方法除了考虑口袋的理化性质外, 还加入了口袋的二级结构信息、深度指数(DPX)和突出指数(CX)特征。另外, 为了克服正负样本极度不平衡的问题, 本工作使用欠采样方法来平衡训练数据集。在独立测试集上, AllosEC在多个评价指标上优于现有的其他方法, SEN、SPE、PRE和MCC分别为0.708、0.915、0.405和0.486。这样, 本工作提供了性能良好的蛋白质变构位点预测方法AllosEC。

关键词

蛋白质变构, 理化性质, 二级结构, 集成模型

An Ensemble Model for Protein Allosteric Site Prediction

Shijie Qiao, Fangrui Hu, Chunhua Li*

College of Chemistry and Life Science, Beijing University of Technology, Beijing

Received: Aug. 22nd, 2024; accepted: Sep. 15th, 2024; published: Sep. 24th, 2024

Abstract

Allostery is an important mechanism for regulating protein functions, which is essential for many biological processes. Compared with orthosteric regulators, allosteric regulators have higher specificity and lower toxicities, which makes allosteric drug design have more advantages than or-

*通讯作者。

thosteric drug design. The discovery of allosteric sites is a prerequisite for allosteric drug design. Currently, experimentally obtained allosteric sites are mostly obtained by chance, and therefore there is an urgent need to develop effective theoretical methods to predict protein allosteric sites. Here, we present an ensemble machine learning method AllosEC for protein allosteric pocket prediction, where besides the pockets' physicochemical properties, their secondary structure information, depth indexes (DPXes) and protrusion indexes (CXes) are considered. In order to overcome the problem of extreme imbalance between positive and negative samples, this work uses an under sampling method to balance the training dataset. AllosEC outperforms other existing methods in multiple evaluation metrics on the independent test set, with SEN, SPE, PRE and MCC of 0.708, 0.915, 0.405 and 0.486, respectively. Thus, this work provides a good method AllosEC for protein allosteric site prediction.

Keywords

Protein Allostery, Physicochemical Properties, Secondary Structure, Ensemble Model

Copyright © 2023 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

变构是蛋白质发挥功能的重要调控机制，它通常是指变构调节剂结合于蛋白质活性位点(正构位点)以外的其它位点(变构位点)，引起蛋白质构象或动力学性质的改变，进而影响蛋白质功能的现象[1] [2] [3]。相比于正构位点，变构位点的结构更具多样性，这使得变构调节剂具有更高的选择性和更少的毒副作用，这些优势使变构调节剂的设计在药物研发中具有重要意义[4] [5] [6]。变构位点的发现是变构药物设计的前提，准确识别蛋白质变构位点成为结构生物学家和药物设计者普遍关注的问题。

目前，变构位点多是由高通量筛选实验偶然所得[7] [8]，这严重阻碍了变构调节剂的设计与开发[9]，因此，发展理论的方法来准确预测蛋白质变构位点受到了研究人员的广泛关注。自变构概念提出以来，许多理论方法被提出用以研究蛋白质变构效应或识别蛋白质变构位点。其中，基于分子动力学的经典方法有双态 Go 模型[10]、AlloMod [11]和 SPACER [12]等方法。这些方法基于物理学思想，需要对构象系综进行采样分析来寻找潜在的变构位点，这通常需要耗费较长的时间。对此，人们提出了基于正则模分析(Normal Mode Analysis, NMA)和粗粒化弹性网络模型(Elastic Network Model, ENM)的方法来识别蛋白质变构位点[13]。除了以上基于动力学的方法外，研究人员还发展了基于进化分析的方法[14]和基于结构网络的方法[15]，但其预测准确性都不够理想，这可能是由于变构位点的特性比较复杂。对此，人们开始试图发展可以考虑多因素的机器学习方法来预测蛋白质变构位点。2013 年，Huang 等人首次将机器学习方法应用于变构位点预测，提出了 Allosite 方法[16]。该方法使用 Fpocket 程序[17]查找蛋白质表面潜在的变构口袋，并将其理化特性作为口袋的描述符来训练模型。2014 年，Panjkovich 等人利用蛋白质结合虚拟配体产生的柔性变化信息，提出了 PARS [18]方法。2017 年，Song 等人将口袋描述符和蛋白质结合虚拟配体产生的柔性变化信息作为特征，提出了 AllositePro 方法[19]。

本工作中，我们提出了一种集成的机器学习分类模型 AllosEC (Ensemble Classification model for protein allosteric site prediction, AllosEC)来预测蛋白质变构口袋。为克服训练数据集中正负样本极不平衡的问题，使用欠采样方法来平衡训练数据集。特征方面，除考虑口袋的理化性质外，还加入了口袋的二

级结构信息、深度指数和突出指数特征。最后，利用所提取的口袋特征来训练堆栈集成模型(Stacking Ensemble Classifier, SEC)以构建变构口袋预测器。

2. 数据与方法

2.1. 数据集

训练数据集来自 ASBench (<https://mdl.shsmu.edu.cn/asbench/>) 基准数据集[20]。ASBench 是蛋白质变构数据库(ASD, <https://mdl.shsmu.edu.cn/ASD/>) [21]中分辨率优于 3.0 Å、序列一致性小于 30%、且未缺失变构残基的蛋白质所构成的数据集。ASBench 有“核心集”和“核心多样性集”两个数据集，后者是从前者中去冗余后得到的，分别由 235 和 147 个晶体结构组成。本工作使用 ASBench 中的“核心多样性集”作为训练集，使用提出 AllositePro 方法的工作[19]中的 24 个变构蛋白质作为独立测试集，两个数据集间没有冗余性。

2.2. 口袋查找和样本生成

首先，对训练集和测试集中的蛋白质，仅保留包含变构残基的蛋白质链[19]。然后，利用 Fpocket 程序[17] (采用默认参数)查找蛋白质表面的潜在变构口袋，该程序还计算了口袋的 19 种理化性质特征。对这些口袋，定义其残基与变构调节剂接触最多的口袋为正样本，其它为负样本。最终，在训练数据集中共找到 2245 个潜在的变构口袋(147 个正样本，2098 个负样本)，在独立测试集中共找到 319 个潜在的变构口袋(24 个正样本，295 个负样本)。

2.3. 特征提取

除了 Fpocket 生成的口袋的 19 种理化性质特征外，本工作还考虑了口袋的二级结构信息、深度指数(Depth index, DPX)、突出指数(Protrusion index, CX)和溶剂可及性特征，共计 38 种特征。对于口袋的二级结构信息，本工作统计了口袋残基在每种二级结构中的数量，以及该数量占口袋残基数量的百分比。这里，用 DSSP [22]识别蛋白质二级结构类型，二级结构包括： 3_{10} -helix (G)、 α -helix (H)、 π -helix (I)、turn (T)、 β -ladder (E)、 β -bridge (B)、bend (S)和不确定的二级结构(M)。采用 PSAIA [23]来计算残基的深度指数 DPX 和突出指数 CX，它们分别表征了残基被埋藏在蛋白质结构中的深度和残基暴露于溶剂中的程度。这里，定义口袋内所有残基 DPX 和 CX 的均值为口袋的深度指数和突出指数。另外，定义口袋的溶剂可及性为口袋内所有残基的相对溶剂可及表面积(Relative Accessible Surface Area, RASA)之和。

2.4. 平衡训练集和构建堆栈集成模型(SEC)

对于一个变构蛋白，Fpocket 常常可以查找到多个口袋，但其中真正的变构口袋往往很少，这会造成正负样本的极度不平衡，从而引起训练在这样数据集上的模型的过拟合现象，因此，平衡训练数据集是需要的。常用于平衡数据集的方法有欠采样和过采样两种方式。这里，本工作使用欠采样方法来平衡训练集，即从训练集中随机抽取与正样本数量相同的负样本来与正样本组合构成训练集。

对于机器学习模型，采用基于多个基分类器的堆栈集成模型(SEC) [24]来搭建预测方法。SEC 由两层组成：第一层为多个基分类器，包括支持向量机(Support Vector Machine, SVM) [25]、朴素贝叶斯(Naive Bayes, NB) [26]和随机森林(Random Forest, RF) [27]模型；第二层是逻辑回归(Logistic Regression, LR) [28]模型。整个模型使用五折交叉验证在训练数据集上进行训练，在独立测试集上进行其性能的测试。

2.5. 评估指标

模型评估指标包括：敏感性(SEN)、特异性(SPE)、精度(PRE)和马修斯相关系数(MCC)，它们的定义

如下：

$$SEN = \frac{TP}{TP + FN} \tag{1}$$

$$SPE = \frac{TN}{TN + FP} \tag{2}$$

$$PRE = \frac{TP}{TP + FP} \tag{3}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{4}$$

其中，真阳性(*TP*)、假阳性(*FP*)、真阴性(*TN*)和假阴性(*FN*)是通过将每个口袋的预测标签与真实标签进行比较而获得的。本工作还使用接收者操作特性(ROC)曲线下的面积(*AUC*)来评估模型的性能。

3. 结果与讨论

3.1. 不同机器学习算法的性能比较

为了选择合适的基分类器来构建模型，本工作在训练集上使用五折交叉验证对 7 个经典分类器进行了性能比较，结果见表 1。与分类器 *k* 近邻(*k*-Nearest Neighbor, *k*NN) [29]、自适应提升树(Adaptive Boosting, Adaboost) [30]、极限梯度提升树(eXtreme Gradient Boosting, XGBoost) [31]和梯度提升树(Gradient Tree Boosting, GTB) [32]相比，NB、RF 和 SVM 在 MCC 得分上表现最佳。因此，本工作采用 NB、RF 和 SVM 作为 SEC 的第一层分类器。另外，与单一分类器 NB、RF 和 SVM 相比，基于多个基分类器的 SEC 模型具有更好的性能，其 SEN、SPE、PRE、MCC 和 AUC 分别为 0.653、0.876、0.791、0.550 和 0.847。因此，本工作使用 SEC 作为 AllosEC 的机器学习分类器。

Table 1. Comparison of different classifiers on the training dataset
表 1. 不同分类器在训练数据集上的比较

Classifier	SEN	SPE	PRE	MCC	AUC
RF	0.673 ± 0.047	0.842 ± 0.044	0.754 ± 0.044	0.528 ± 0.040	0.843 ± 0.024
kNN	0.613 ± 0.042	0.828 ± 0.037	0.717 ± 0.042	0.455 ± 0.041	0.766 ± 0.029
SVM	0.633 ± 0.033	0.880 ± 0.040	0.792 ± 0.053	0.538 ± 0.052	0.840 ± 0.030
Adaboost	0.672 ± 0.080	0.803 ± 0.053	0.709 ± 0.064	0.482 ± 0.100	0.807 ± 0.054
GTB	0.666 ± 0.043	0.813 ± 0.071	0.725 ± 0.078	0.490 ± 0.077	0.791 ± 0.028
XGBoost	0.701 ± 0.041	0.790 ± 0.017	0.701 ± 0.013	0.491 ± 0.030	0.824 ± 0.016
NB	0.523 ± 0.043	0.928 ± 0.026	0.841 ± 0.052	0.509 ± 0.043	0.847 ± 0.031
SEC	0.653 ± 0.017	0.876 ± 0.038	0.791 ± 0.049	0.550 ± 0.041	0.847 ± 0.017

3.2. AllosEC 与现有方法比较

本工作在独立测试集上对 AllosEC 的预测性能进行了评估，结果见表 2。为了将 AllosEC 与其他方法进行比较，表 2 也列出了 PARS、Allosite 和 AllositePro 方法的相应结果(来自 AllositePro 方法[19]的文章)。从表 2 可见，AllosEC 模型的 SEN、SPE、PRE 和 MCC 分别为 0.708、0.915、0.405 和 0.486，优于其他方法。从 MCC 来看，AllosEC 比 AllositePro 性能提高了 0.098，比 PARS 和 Allosite 性能提高更显著，分

别为 0.387 和 0.222。分析模型性能提升的可能原因，主要有两方面：特征方面，除了口袋的理化性质外，AllosEC 还考虑了口袋的二级结构信息、深度指数和突出指数；模型方面，SEC 集成了几个基分类器的结果，泛化性能得到了一定的提升。

Table 2. Comparison of AllosEC with other methods on independent test set
表 2. AllosEC 在独立测试集上与其他方法比较

Methods	SEN	SPE	PRE	MCC	AUC
PARS	0.375	0.756	0.180	0.099	-
Allosite	0.500	0.858	0.245	0.264	-
AllositePro	0.625	0.885	0.333	0.388	-
AllosEC	0.708	0.915	0.405	0.486	0.869

-: AllositePro 方法的文章中未提供 AUC 数据。

3.3. 案例分析

作为研究案例，图 1 显示了 AllosEC 对两种蛋白质变构口袋的预测结果。第一种是 CC 趋化因子受体(PDB ID:5C4T) [33]，实验报道了一个由 13 个残基(Q329, I328, T325, A321, L483, W317, A497, F498, A496, L353, F506, Y502, L505)构成的变构口袋。图 1(a)显示，AllosEC 预测出两个口袋为阳性，其一包含 24 个残基(F506, V494, T325, Q484, L483, Q487, L505, A496, Y502, A497, F498, V480, M358, L324, I328, K354, L353, A357, L501, Q329, Q495, A321, W317, I492)，完全覆盖了实验报道的变构残基；尽管，另一个预测为阳性的口袋未覆盖目前已知的变构残基，但其可能是潜在的变构口袋，这需要未来的实验来证实。

第二种是嗜热球菌 OT3 的生物素蛋白连接酶(PDB ID:1WQW) [34]，实验报道了一个由 42 个残基(N134, G139, Q42, T141, L60, G47, V115, L62, V136, T20, G127, G114, R51, W53, P137, A140, S55, G129, R48, G45, H46, W61, I112, M44, L130, K111, D104, N103, P56, A113, L49, S63, E54, I128, E24, K52, T22, S21, N131, N23, K135, L116)构成的变构口袋。图 1(b)显示，AllosEC 准确预测出 1 个变构口袋，其中包含 52 个残基(V115, N103, L116, Y227, R48, G228, V230, P102, W101, I226, G114, G127, H46, G129, W53, Q42, S21, N23, G45, W61, I128, L62, T22, G47, K111, S63, G196, L195, D229, S231, R233, R51, G198, P137, A140, N131, E54, K52, N134, K135, V136, P56, A113, N138, G139, D104, K100, L232, L213, I194, V225, V192)，覆盖了实验报道的 33 个变构残基。

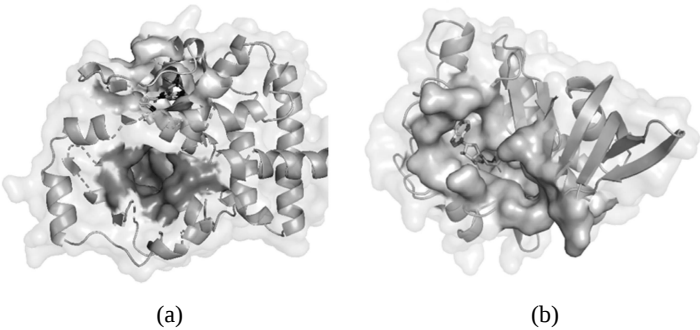


Figure 1. Predicted results of AllosEC for allosteric pockets of two allosteric proteins with PDBIDs of 5C4T (a) and 1WQW (b), respectively. Protein and allosteric regulators are shown in cartoon and stick models respectively. The true and potential allosteric pockets are shown in light and dark colors respectively

图 1. AllosEC 对两种蛋白质变构口袋的预测结果，其 PDBID 分别为 5C4T (a)和 1WQW (b)。蛋白质和变构调节剂分别用卡通和棍棒模型表示。真实和潜在的变构口袋分别显示为浅色和深色

4. 结论

有效预测蛋白质变构位点对理解蛋白质变构机制和基于结构的变构药物设计至关重要。本工作提出了一种基于多个基分类器的集成模型 AllosEC 来预测蛋白质变构口袋。该方法采用欠采样技术来平衡训练集数据,以避免模型的过拟合。除了口袋的理化特性,方法还考虑了口袋的二级结构信息、深度和突出指数。在独立测试集上,AllosEC 的 SEN 和 MCC 分别为 0.708 和 0.486,优于其它方法,显示了对蛋白质变构位点预测好的性能。

目前,缺乏实验确定的蛋白质变构位点数据仍然是阻碍预测方法发展的主要原因之一。尽管目前考虑多因素的机器学习算法有好的鲁棒性,但新的有效的特征仍有待开发以进一步提高预测准确性。可能下述特征的考虑有助于预测性能的进一步提升。一是动力学信息,蛋白质变构需要通过残基涨落动力学的偶合来改变正构位点的功能,因此动力学性质的考虑是有益的;二是正构位点信息,蛋白质变构位点与正构位点具有一定的关系,考虑正构位点信息也是有益的。随着人们对有益特征的探索,相信基于机器学习算法的蛋白质变构位点预测的准确性会不断提高。

基金项目

国家自然科学基金项目(32271294, 31971180)。

参考文献

- [1] Greener, J.G. and Sternberg, M.J. (2018) Structure-Based Prediction of Protein Allostery. *Current Opinion in Structural Biology*, **50**, 1-8. <https://doi.org/10.1016/j.sbi.2017.10.002>
- [2] Liu, J. and Nussinov, R. (2016) Allostery: An Overview of Its History, Concepts, Methods, and Applications. *PLOS Computational Biology*, **12**, e1004966. <https://doi.org/10.1371/journal.pcbi.1004966>
- [3] Zha, J., Li, M., Kong, R., et al. (2022) Explaining and Predicting Allostery with Allosteric Database and Modern Analytical Techniques. *Journal of Molecular Biology*, **434**, Article ID: 167481. <https://doi.org/10.1016/j.jmb.2022.167481>
- [4] Lu, S., He, X., Ni, D., et al. (2019) Allosteric Modulator Discovery: From Serendipity to Structure-Based Design. *Journal of Medicinal Chemistry*, **62**, 6405-6421. <https://doi.org/10.1021/acs.jmedchem.8b01749>
- [5] Guarnera, E. and Berezovsky, I.N. (2016) Allosteric Sites: Remote Control in Regulation of Protein Activity. *Current Opinion in Structural Biology*, **37**, 1-8. <https://doi.org/10.1016/j.sbi.2015.10.004>
- [6] Cheng, X. and Jiang, H. (2019) Allostery in Drug Development. In: Zhang, J. and Nussinov, R., Eds., *Protein Allostery in Drug Discovery*, Advances in Experimental Medicine and Biology, Vol. 1163, Springer, Berlin, 1-23. https://doi.org/10.1007/978-981-13-8719-7_1
- [7] Jiang, Y. and Kalodimos, C.G. (2017) NMR Studies of Large Proteins. *Journal of Molecular Biology*, **429**, 2667-2676. <https://doi.org/10.1016/j.jmb.2017.07.007>
- [8] Xiao, S., Verkhivker, G.M. and Tao, P. (2022) Machine Learning and Protein Allostery. *Trends in Biochemical Sciences*, **48**, 375-390. <https://doi.org/10.1016/j.tibs.2022.12.001>
- [9] Gulati, S., Palczewski, K., Engel, A., et al. (2019) Cryo-EM Structure of Phosphodiesterase 6 Reveals Insights into the Allosteric Regulation of Type I Phosphodiesterases. *Science Advances*, **5**, v4322. <https://doi.org/10.1126/sciadv.aav4322>
- [10] Qi, Y., Wang, Q., Tang, B., et al. (2012) Identifying Allosteric Binding Sites in Proteins with a Two-State Go Model for Novel Allosteric Effector Discovery. *Journal of Chemical Theory and Computation*, **8**, 2962-2971. <https://doi.org/10.1021/ct300395h>
- [11] Weinkam, P., Pons, J. and Sali, A. (2012) Structure-Based Model of Allostery Predicts Coupling between Distant Sites. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 4875-4880. <https://doi.org/10.1073/pnas.1116274109>
- [12] Goncarenco, A., Mitternacht, S., Yong, T., et al. (2013) Spacer: Server for Predicting Allosteric Communication and Effects of Regulation. *Nucleic Acids Research*, **41**, W266-W272. <https://doi.org/10.1093/nar/gkt460>
- [13] Ma, X., Meng, H. and Lai, L. (2016) Motions of Allosteric and Orthosteric Ligand-Binding Sites in Proteins Are Highly Correlated. *Journal of Chemical Information and Modeling*, **56**, 1725-1733.

- <https://doi.org/10.1021/acs.jcim.6b00039>
- [14] Suel, G.M., Lockless, S.W., Wall, M.A., *et al.* (2003) Evolutionarily Conserved Networks of Residues Mediate Allosteric Communication in Proteins. *Nature Structural Biology*, **10**, 59-69. <https://doi.org/10.1038/nsb881>
 - [15] Wang, J., Jain, A., McDonald, L.R., *et al.* (2020) Mapping Allosteric Communications within Individual Proteins. *Nature Communications*, **11**, Article No. 3862. <https://doi.org/10.1038/s41467-020-17618-2>
 - [16] Huang, W., Lu, S., Huang, Z., *et al.* (2013) Allosite: A Method for Predicting Allosteric Sites. *Bioinformatics*, **29**, 2357-2359. <https://doi.org/10.1093/bioinformatics/btt399>
 - [17] Le Guilloux, V., Schmidtke, P. and Tuffery, P. (2009) Fpocket: An Open Source Platform for Ligand Pocket Detection. *BMC Bioinformatics*, **10**, Article No. 168. <https://doi.org/10.1186/1471-2105-10-168>
 - [18] Panjkovich, A. and Daura, X. (2014) Pars: A Web Server for the Prediction of Protein Allosteric and Regulatory Sites. *Bioinformatics*, **30**, 1314-1315. <https://doi.org/10.1093/bioinformatics/btu002>
 - [19] Song, K., Liu, X., Huang, W., *et al.* (2017) Improved Method for the Identification and Validation of Allosteric Sites. *Journal of Chemical Information and Modeling*, **57**, 2358-2363. <https://doi.org/10.1021/acs.jcim.7b00014>
 - [20] Huang, W., Wang, G., Shen, Q., *et al.* (2015) ASBench: Benchmarking Sets for Allosteric Discovery. *Bioinformatics*, **31**, 2598-2600. <https://doi.org/10.1093/bioinformatics/btv169>
 - [21] Shen, Q., Wang, G., Li, S., *et al.* (2016) Asd v3.0: Unraveling Allosteric Regulation with Structural Mechanisms and Biological Networks. *Nucleic Acids Research*, **44**, D527-D535. <https://doi.org/10.1093/nar/gkv902>
 - [22] Kabsch, W. and Sander, C. (1983) Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers*, **22**, 2577-2637. <https://doi.org/10.1002/bip.360221211>
 - [23] Mihel, J., Sikic, M., Tomic, S., *et al.* (2008) Psai-Protein Structure and Interaction Analyzer. *BMC Structural Biology*, **8**, Article No. 21. <https://doi.org/10.1186/1472-6807-8-21>
 - [24] Wolpert, D.H. (1992) Stacked Generalization. *Neural Networks*, **5**, 241-259. [https://doi.org/10.1016/S0893-6080\(05\)80023-1](https://doi.org/10.1016/S0893-6080(05)80023-1)
 - [25] Cherkassky, V. (1997) The Nature of Statistical Learning Theory. *IEEE Transactions on Neural Networks*, **8**, 1564. <https://doi.org/10.1109/TNN.1997.641482>
 - [26] Zhang, H. (2004) The Optimality of Naive Bayes. *Proceedings FLAIRS*, **2**, 562-567.
 - [27] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. <https://doi.org/10.1023/A:1010933404324>
 - [28] Kleinbaum, D.G. and Klein, M. (2010) Logistic Regression. Springer, New York. <https://doi.org/10.1007/978-1-4419-1742-3>
 - [29] Cover, T.M.T. (1968) Rates of Convergence for Nearest Neighbor Procedures. *Proceedings of the Hawaii International Conference on System Sciences*, Honolulu, 29-30 January 1968, 413-415.
 - [30] Freund, Y. and Schapire, R.E. (1997) A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, **55**, 119-139. <https://doi.org/10.1006/jcss.1997.1504>
 - [31] Chen, T. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 785-794. <https://doi.org/10.1145/2939672.2939785>
 - [32] Friedman, J.H. (2002) Stochastic Gradient Boosting. *Computational Statistics & Data Analysis*, **38**, 367-378. [https://doi.org/10.1016/S0167-9473\(01\)00065-2](https://doi.org/10.1016/S0167-9473(01)00065-2)
 - [33] Scheepstra, M., Leysen, S., van Almen, G.C., *et al.* (2015) Identification of an Allosteric Binding Site for Rorgammat Inhibition. *Nature Communications*, **6**, Article No. 8833. <https://doi.org/10.1038/ncomms9833>
 - [34] Bagautdinov, B., Kuroishi, C., Sugahara, M., *et al.* (2005) Crystal Structures of Biotin Protein Ligase from *Pyrococcus horikoshii* ot3 and Its Complexes: Structural Basis of Biotin Activation. *Journal of Molecular Biology*, **353**, 322-333. <https://doi.org/10.1016/j.jmb.2005.08.032>