REFERENCES

[1] W, Shoombuatong., et al Prediction of human leukocyte antigen gene using *k*-nearest neighbour classifier based on spectrum kernel. *ScienceAsia.* 39:42–49, 2013.

[2] W, Shoombuatong., et al HIV-1 CRF01_AE coreceptor usage prediction using kernel methods based logistic model trees. *Computers in Biology and Medicine*, 42: 885–889, 2012.

[3] W, Shoombuatong., et al. Predicting Protein Crystallization Using a Simple Scoring Card Method, *2013 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, 23-30, 2013.

[4] C, Wedekind., et al The major histocompatibility complex and perfumers descriptions of human body odors. *Evol Psychol.* 5: 330–43, 2007.

[5] S, Nail., The human HLA system. *J Indian Rheumatol Assoc*. 11: 79–83, 2003.

[6] U, Shankarkumar., The human leukocyte antigen (HLA) system. *Int J Hum Genet*. 4: 91–103, 2004.

[7] V, Apanius., The nature of selection on the major histocompatibility complex. *Crit Rev Immunol*. 17: 179–224, 1997.

[8] A, Browning., HLA and MHC genes, molecules and function, Bios Scientific Publishers, Oxford, 1996.

[9] Y, Feng., et al. HIV-1 entry cofactor: functional cDNA cloning of a seven-transmembrane, G-protein-coupled receptor, *Science.* 272 872–877, 1996.

[10] E.A. Berger., et al. A new classification for HIV-1. *Nature.* 391 (15), 240, 1998.

[11] M. Koot., et al. Progmosis value of HIV-1 syncytium-inducing phenotype for rate of CD4+ cell depletion and progression to AIDs. *Ann. Intern. Med*. 118:681–688, 1993.

[12] D, Richman., et al. The impact of the syncytium-inducing phenotype of HIV on disease progression. *J. Infect. Dis*. 169: 968–974, 1994.

[13] M, Norin., et al. Protein models in drug discovery. *Curr Opin Drug Discov Devel*, 4(3): 284-90, 2001.

[14] HM, Berman., et al, The Protein Data Bank and the challenge of structural genomics *Nature structural biology*, 7:957-959, 2000.

[15] R, Hui., et al. High-throughput protein crystallization. *Journal of structural biology*, 142(1): 154-161, 2003.

[16] Z, Xing, et al. A brief survey on sequence classification. ACM SIGKDD Explorations Newsletter, 12(1): 40-48, 2010

[17] N. Lesh., et al Mining features for sequence classi cation. *In KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, 342-346, 1999.

[18] C. S, Leslie., et al. The spectrum kernel: A string kernel for SVM protein classification. *In Pacific Symposium on Biocomputing*, 566-575, 2002.

[19] C.S, Leslie., et al. Fast string kernels using inexact matching for protein sequences. *Journal of Machine Learning Research*, 5:1435-1455, 2004.

[20] M, Deshpande., et al. *Evaluation of techniques for classifying biological sequences*. *In PAKDD '02: Proceedings of the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, 417-431, 2002.

[21] G, Dong., et al. Sequence Data Mining, Springer US, 47-65., 2007.

[22] N. A. Chuzhanova., et al. Feature selection for genetic sequence classification. *Bioinformatics*, 14(2):139-143, 1998.

[23] E, Keogh., et al. On the need for time series data mining benchmarks: a survey and empirical demonstration. *In KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 102-111, 2002.

[24] D. D. Lewis., Naive (bayes) at forty: The independence assumption in information retrieval. *In ECML' 98: The 10th European Conference on Machine Learning*, 4-15, 1998.

[25] S.-B. Kim., et al. Some effiective techniques for naive bayes text classification. *IEEE Transactions on Knowledge and Data Engineering*, 18(11):1457-1466, 2006.

[26] B. Cheng., et al. Protein classification based on text document classification techniques. *Proteins*, 1(58):855-970, 2005.

[27] R. Durbin., et al. Chapter 3. Markov Chain and Hidden Markov Model. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, Cambridge University Press, 47-65, 1998.

[28] O. Yakhnenko., et al. Discriminatively trained markov model for sequence classification. *In ICDM '05: Proceedings of the Fifth IEEE International Conference on Data Mining*, 498-505, 2005.

[29] P. K. Srivastava, et al. HMM-ModE-Improved classifiation using profile hidden Markov models by optimising the discrimination threshold and modifying emission probabilities with negative training sequences. *BMC Bioinformatics*, 8(104), 2007.

[30] V, Vapnik., The Nature of Statistical Learning Theory, Springer, 1995.

[31] XL, Xia., et al. Two-stage gene selection for support vector machine classification of microarray data. *Int. J. Modelling, Identification and Control*, 8: 164-171, 2009.

[32] C, Park., et al. Classification of Gene Functions Using Support Vector Machine for Time-Course Gene Expression Data. *Computational Statistics and Data Analysis*, 52 (5): 2578–2587, 2008.

[33] T, Sing et al, Predicting HIV co-receptor usage based on genetic and clinical covariates. *Antiviral Therapy*, 12: 1097-1106, 2007.

[34] C, To., et al. A combination of kernel methods and genetic programming for gene expression pattern classification. *Research, Innovation and Vision for the Future*, 214-221, 2006.

[35] C. Cortes., et al. Support-vector networks. *Mach.Learning*, 20: 273–297, 1995.

[36] V, Vapnik, Statistical Learning Theory, *John Wiley & Sons*, 1998

[37] Gartner T., A survey of kernels for structured data. *SIGKDD Explor Newslett*, 5:49–58, 2003.

[38] C, Watkins., Kernel from matching operation, Tech. rep. Department of Computer Science, Royal Holloway, Univ of London, 1990.

[39] H, Lodhi H., et al. Text classification using string kernels. *J Mach Learn Res*, 2: 419–44, 2003.

[40] Leslie C., et al Mismatch string kernels for discriminative protein classification. *Bioinformatics*, 20(4):467-76, 2004.

[41] C, Saunders., et al.Syllables and other string kernel extensions. *Proceedings of the 19th International Conference on Machine Learning (ICML02)*, 530–7, 2002.

[42] S, Sonnenburg., et al. Learning interpretable SVMs for biological sequence classification. *In RECOMB '05: The Ninth Annual International Conference on Research in Computational Molecular Biology*, 389-407, 2005.

[43] H, Saigo., et al. Protein homology detection using string alignment kernels. *Bioinformatics*, 20(11):1682-1689, 2004.

[44] Y. Saeys., et al. A review of feature selection techniques in bioinformatics. *Bioinformatics* 23(19), 2507-2517, 2007.

[45] M, Ben-Bassat., Pattern recognition and reduction of dimensionality. In Krishnaiah,P. and Kanal,L., (eds.) *Handbook of Statistics II, Vol. 1. North-Holland, Amsterdam,* 773–791, 1982.

[46] J, Holland., Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor, 1975.

[47] J, Kittler., Pattern Recognition and Signal Processing, Chapter Feature Set Search Algorithms Sijthoff and Noordhoff, Alphen aan den Rijn, Netherlands, 41–60, 1978.

[48] P, Duda., et al. (2001) Pattern Classification. Wiley, New York.

[49] N. Landweh., et al Logistic Model Trees. Machine Learning, 59: 161-205, 2005.

[50] L. Breiman., Random forest *Machine Learning*, 45: 5-32

[51] H-L, Huang., et al. Prediction and analysis of protein solubility using a novel scoring card method with dipeptide composition, *BMC Bioinformatics* , 13:(Suppl 17):S3, 2012.

[52] S-Y. Ho., et al. Intelligent Evolutionary Algorithms for Large Parameter Optimization Problems. *IEEE Transactions on Evolutionary Computation*, 8(6): 522-541, 2004.

[53] AP. Bradley., The use of the area under the ROC curve in the evaluation of machine learning algorithms, *Pattern Recognition*, 30: 1145-1159, 1996.

[54] J, Friedman., et al Additive logistic regression: a statistical view of boosting. *The Annals of Statistic*, 38(2), 337–374, 2000.

[55] R. Quinlan., C4.5: Programs for Machine Learning. Morgan Kaufmann, Morgan Kaufmann Publishers Inc. San Francisco, 1993.

[56] L, Breiman et al. Classification and Regression Trees. Wadsworth, 1984.

[57] MJ, Ma., et al. Gene classification using codon usage and support vector machines. *IEEE ACM Trans Comput Biol Bioinformatics*, 6: 134–43, 2009

[58] JW, Han., et al. Data Mining: Concepts and Techniques. Academic Press, 2001.

[59] D, Aha., et al. Instance-based learning algorithms. *Mach Learn*, 6: 37–66, 1991.

[60] MN, Nguyen., et al. Di-codon usage for gene classification. *Lect Notes Comput Sci*, 5780: 211–21, 2009.

[61] J, Robinson et al. IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. *Nucleic Acids Res*, 31: 311–4, 2003.

[62] J, Robinson et al. IMGT/HLA and IMGT/MHC: sequence databases for the human major histocompatibility complex. *Tissue Antigens*, 55: 280–287, 2000.

[63] J, Robinson et al. IMGT/HLA and IMGT/MHC: sequence databases for the human major histocompatibility complex. *Nucleic Acids Res*, 29: 210–3, 2001

[64] J, Robinson et al. The IMGT/HLA database. Nucleic Acids Res 37, 1013–7, 2009.

[65] SG, Marsh., et al. Nomenclature for factors of the HLA system. Tissue Antigens, 57, 236–83, 2001.

[66] M. Kober., stringkernels (String Kernel Methods for kernlab); software available at http://cran.r-project.org/web/packages/stringkernels, 2010.

[67] A, Karatzoglou., kernlab (Kernel-based Machine Learning Lab); software available at http://cran.r-project.org/web/packages/kernlab, 2009.

[68] AD, Brooks., knnflex' (knnflex: A more flexible KNN); software available at http://cran.r-project.org/web/packages/knnflex, 2009.

[69] B, Ripley B et al., nnet' (Feed- forward Neural Networks and Multinomial Log-Linear Models); software available at http://cran.r-project.org/ web/packages/nnet, 2009.

[70] JL, Milhon., et al. Updated codon usage in Schistosoma. *Exp Parasitol*, 80: 353–6, 1995.

[71] M, Mitreva., et al Codon usage patterns in Nematoda: analysis based on over 25 million codons in thirty-two species. *Genome Biol*, 7: R75, 2006.

[72] KD, Pruitt et al. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res* , 37 (Database issue), D32–6 2009.

[73] J, Weber., et al. HIV type 1 tropism and inhibitors of viral entry: clinical implications, *AIDS Rev*. 8 60–77, 2006.

[74] S.S. Hwang., et al. Identification of the envelope V3 loop as the primary determinant of cell tropism in HIV-1, *Science*, 253:71–74, 1991.

[75] S. Pillai A new perspective on V3 phenotype prediction, *AIDS Res. Hum. Retrov*. 19: 145–149, 2003.

[76] W, Resch et al. Improved success of phenotype prediction of the human immunodeficiency virus type 1 from envelop variable loop 3 sequence using neural networks, *Virology*, 288:51–62, 2001.

[77] M.A., Jensen., et al. Improved coreceptor usage prediction and genotypic monitoring of R5-to-X4 transition by motif analysis of human immunodeficiency virus type 1 env V3 loop sequences, *J. Virol*, 77:13376–13388, 2003.

[78] R.A. Fouchier.,et al. Phenotype-associated sequence variation in the third variable of the human immunodeficiency virus type1 gp120 molecule, *J. Virol*. 66:3183–3187, 1992

[79] P. Delobel., et al. Population-based sequencing of the V3 region of env for predicting the coreceptor usage of human immunodeficiency virus type 1 quasispecies, *J. Clin. Microbio*l. 45 (5): 1572–1580, 2007.

[80] L.P. Vandekerckhove ., et al. European consensus group on clinical management of tropism testing. European guidelines on clinical management of HIV-1 tropism testing, *Lancet Infect. Dis*, 11 (5): 394–407, 2011.

[81] C. Garrido, et al., Evaluation of eight different bioinformatics tools to predict viral tropism in different human immunodeficiency virus type1 subtypes, *J. Clin. Microbiol*. 46: 887–891, 2008.

[82] M.A. Jensen., et al. A reliable phenotype predictor for human immunodeficiency virus type 1 subtype C based on envelope V3 sequences. *J Virol*. 80(10):4698-704, 2006.

[83] O Sander., et al. Structural Descriptors of gp120 V3 Loop for the Prediction of HIV-1 Coreceptor Usage, PLoS Comput Biol, 3(3), 2007

[84] S.L. Lamers., et al Prediction of R5, X4, and R5X4 HIV-1 Coreceptor Usage with Evolved Neural Networks. *EEE/ACM Trans Comput Biol Bioinform*. 5(2): 291–300, 2008.

[85] S, Boisvert et al**.** HIV-1 coreceptor usage prediction without multiple alignments: an application of string kernels, *Retrovirology* 2008, 5:110, 2008

[86] K. Hornik., et al. LMT (logistic model trees implement), software, available from: /http:// cran.r-project.org/web/packages/RWeka/index.html, 2011.

[87] T. Sing., et al. Predicting HIV coreceptor usage based on genetic and clinical covariates, *Antivir. Ther*. 12:1097–1106, 2007.

[88] M, Norin., et al. Protein models in drug discovery. Curr Opin Drug Discov Devel 4: 284-290, 2001.

[89] HM, Berman et al. The Protein Data Bank and the challenge of structural genomics. Nature Structural Biology 7: 957-959, 2000.

[90] R, Hui., et al. High-throughput protein crystallization. J Struct Biol 142: 154-161, 2003.

[91] JM, Canaves., et al. Protein biophysical properties that correlate with crystallization success in Thermotoga maritima: Maximum clustering strategy for structural genomics. Journal of Molecular Biology 344: 977-991, 2004.

[92] CS, Goh., et al. Mining the structural genomics pipeline: Identification of protein properties that affect high-throughput experimental analysis. *Journal of Molecular Biology*, 336: 115-130, 2004.

[93] IM, Overton., et al. (2006) A normalised scale for structural genomics target ranking: the OB-Score. *FEBS Lett* 580: 4005-4009, 2006.

[94] P, Smialowski., et al Will my protein crystallize? A sequence-based predictor. *Proteins-Structure Function and Bioinformatics* 62: 343-355, 2006.

[95] K, Chen K., et al. Prediction of protein crystallization using collocation of amino acid pairs. *Biochem Biophys Res Commun* 355: 764-769, 2007.

[96] L, Slabinski., et al. XtalPred: a web server for prediction of protein crystallizability. *Bioinformatics* 23: 3403-3405, 2007.

[97] IM, Overton., et al. ParCrys: a Parzen window density estimation approach to protein crystallization propensity prediction. *Bioinformatics* 24: 901-907, 2008.

[98] L, Kurgan., et al. CRYSTALP2: sequence-based protein crystallization propensity prediction. *BMC Struct Biol* 9: 50, 2009.

[99] KK, Kandaswamy., et al. SVMCRYS: an SVM approach for the prediction of protein crystallization propensity from protein sequence. *Protein Pept Lett* 17: 423-430, 2010.

[100] MJ, Mizianty .,et al Sequence-based prediction of protein crystallization, purification and production propensity. *Bioinformatics* 27: i24-33, 2011.

[101] S, Jahandideh ., et al. RFCRYS: Sequence-based protein crystallization propensity prediction by means of random forest. *Journal of Theoretical Biology* 306: 115-119, 2012.

[102] P, Charoenkwan et al Predicting protein crystallization using an ensemble scoring card method with estimating propensity scores of p-collocated amino acid pairs, PlosOne. (Accepted)

[103] Derewenda ZS Application of protein engineering to enhance crystallizability and improve crystal properties. Acta Crystallographica Section D-Biological Crystallography 66: 604-615, 2010.

[104] Pickett SD et al Empirical Scale of Side-Chain Conformational Entropy in Protein-Folding. Journal of Molecular Biology 231: 825-839, 1993.

[105] Goldschmidt L et al Toward rational protein crystallization: A Web server for the design of crystallizable protein variants. Protein Science 16: 1569-1576, 2007.

[106] Cooper DR et al. Protein crystallization by surface entropy reduction: optimization of the SER strategy. Acta Crystallographica Section D-Biological Crystallography 63: 636-645, 2007

[107] Longenecker KL et al Protein crystallization by rational mutagenesis of surface residues: Lys to Ala mutations promote crystallization of RhoGDI. Acta Crystallographica Section D-Biological Crystallography 57: 679-688, 2001.

[108] Birtley JR et al Crystallization of foot-and-mouth disease virus 3C protease: surface mutagenesis and a novel crystal-optimization strategy. Acta Crystallographica Section D-Biological Crystallography 61: 646-650, 2005.

[109] MartinezHackert E et al Crystallization, X-ray studies, and site-directed cysteine mutagenesis of the DNA-binding domain of OmpR. Protein Science 5: 1429-1433, 1996.

[110] Mateja A et al. The impact of Glu -> Ala and Glu -> Asp mutations on the crystallization properties of RhoGDI: the structure of RhoGDI at 1.3 angstrom resolution. Acta Crystallographica Section D-Biological Crystallography 58: 1983-1991, 2002.

[111] Garrard SM et al Expression, purification, and crystallization of the RGS-like domain from the rho nucleotide exchange factor, PDZ-RhoGEF, using the surface entropy reduction approach. Protein Expression and Purification 21: 412-416, 2001.

[112] Janda I et al. Harvesting the high-hanging fruit: the structure of the YdeN gene product from Bacillus subtilis at 1.8 angstrom resolution. Acta Crystallographica Section D-Biological Crystallography 60: 1101-1107, 2004.

[113] Munshi S et al Structure of apo, unactivated insulin-like growth factor-1 receptor kinase at 1.5 angstrom resolution. Acta Crystallographica Section D-Biological Crystallography 59: 1725-1730, 2003.

[114] Bielnicki J et al. B-subtilis ykuD protein at 2.0 A resolution: Insights into the structure and function of a novel, ubiquitous family of bacterial enzymes. Proteins-Structure Function and Bioinformatics 62: 144-151, 2006.

[115] Yip CK et al. Structural characterization of the molecular platform for type III secretion system assembly. Nature 435: 702-707, 2005.

[116] Devedjiev Y et al. The structure and ligand binding properties of the B. subtilis YkoF gene product, a member of a novel family of thiamin/HMP-binding proteins. Journal of Molecular Biology 343: 395-406, 2004.

[117] Boeshans KM et al. Purification, crystallization and preliminary X-ray diffraction analysis of the phage T4 vertex protein gp24 and its mutant forms. Protein Expr Purif 49: 235-243, 2006.

[118] Guo Y et al A single point mutation changes the crystallization behavior of Mycoplasma arthritidis-derived mitogen. Acta Crystallographica Section F-Structural Biology and Crystallization Communications 62: 238-241, 2006.

[119] Honjo E et al. Mutagenesis of the crystal contact of acidic fibroblast growth factor. Journal of Synchrotron Radiation 15: 285-287, 2008.

[120] Al-Ayyoubi M et al Crystal structure of human maspin, a serpin with antitumor properties - Reactive center loop of maspin is exposed but constrained. Journal of Biological Chemistry 279: 55540-55544, 2004.

[121] Schwede TF et al Homogenization and crystallization of histidine ammonia-lyase by exchange of a surface cysteine residue. Protein Engineering 12: 151-153, 1990.

[122] Patel SB et al. Lattice stabilization and enhanced diffraction in human p38 alpha crystals by protein engineering. Biochimica Et Biophysica Acta-Proteins and Proteomics 1696: 67-73, 2004.

[123] Nickerson D et al An improved procedure for the preparation of X-ray diffraction quality crystals of cytochrome P450(cam). Acta Crystallographica Section D-Biological Crystallography 54: 470-472, 1998.

[124] Hibi T et al Escherichia coli B gamma-glutamylcysteine synthetase: modification, purification, crystallization and preliminary crystallographic analysis. Acta Crystallographica Section D-Biological Crystallography 58: 316-318, 2002.

[125] Kessler D et al Structure and action of urocanase. Journal of Molecular Biology 342: 183-194, 2004.

[126] Gustin SE et al. Expression, crystallization and derivatization of the complete extracellular domain of the beta(c) subunit of the human IL-5, IL-3 and GM-CSF receptors. European Journal of Biochemistry 268: 2905-2911, 2001.

[127] Klein C et al Engineering a Heavy-Atom Derivative for the X-Ray Structure-Analysis of Cyclodextrin Glycosyltransferase. Protein Engineering 4: 65-67, 1990.

[128] Adams MJ et al Site-Directed Mutagenesis to Facilitate X-Ray Structural Studies of Leuconostoc-Mesenteroides Glucose-6-Phosphate-Dehydrogenase. Protein Science 2: 859-862, 1993.