In silico design of pharmaceutical compounds with large-scale kernel methods
An Annotated Bibliography

Alexandre Drouin
Université Laval
alexandre.drouin.8@ulaval.ca

Highly interesting

[1] S. An, W. Liu, and S. Venkatesh, "Fast cross-validation algorithms for least squares support vector machine and kernel ridge regression," *Pattern Recognition*, vol. 40, no. 8, pp. 2154–2162, 2007.

> A fast cross-validation algorithm for Kernel Ridge Regression. The proposed method is shown to greatly outperform the naïve multi-fold cross-validation method. The authors also present a Cholesky decomposition based approximate version of their algorithm for large-scale learning problems. Such an algorithm is highly interesting, since training a pan-specific predictor on large datasets requires cross-validation, which requires a great amount of computational resources.

[2] F. Bach, "Sharp analysis of low-rank kernel matrix approximations," *arXiv preprint arXiv :1208.2015*, 2012.

> A review of kernel matrix approximation methods and an analysis of the predictive performance of predictors trained using these methods. Most of these approximations being based on a random subset of p columns from the kernel matrix, choosing an appropriate value for p can be difficult and requires the use of techniques such as cross-validation. The authors provide a method for selecting p to be linear in term of the number of training examples.

[3] C. Cortes, M. Mohri, and A. Talwalkar, "On the impact of kernel approximation on learning accuracy," in *Conference on Artificial Intelligence and Statistics*, 2010, pp. 113–120.

> A study of the impact of kernel approximation on the learning accuracy of kernel based algorithms. This impact is studied for Support Vector Machines, Kernel Ridge Regression and graph Laplacian-based regularization algorithms. The authors provide bounds that help determine the degree of approximation that can be tolerated in the estimation of the kernel matrix. The Nyström low rank matrix approximation method is presented and its application to solving Kernel Ridge Regression is discussed.

[4] S. Giguère, M. Marchand, F. Laviolette, A. Drouin, and J. Corbeil, "Learning a peptide-protein binding affinity predictor with kernel ridge regression," *arXiv preprint arXiv :1207.7253*, 2012.

> The Generic String kernel is an elegant generalization of eight state-of-the-art string kernels for small biomolecules.

[5] H. Kashima, I. Tsuyoshi, and M. SUGIYAMA, "Recent advances and trends in large-scale kernel methods," *IEICE transactions on information and systems*, vol. 92, no. 7, pp. 1338–1353, 2009.

> An extensive review of methods allowing to scale kernel methods to large-scale learning problems. Although this review is slightly outdated, it presents the basics of many current state-of-the-art methods in scaling kernel methods to large datasets.

[6] S. Kumar, M. Mohri, and A. Talwalkar, “Ensemble nystrom method,” in *Neural Information Processing Systems*, vol. 7, 2009, p. 223.

An interesting extension to the Nyström method based on a weighted combination of Nyström matrix approximations. The authors show that a significant performance increase can be obtained by using the Ensemble Nyström method. Various techniques for selecting the combination weights are discussed and analyzed.

[7] ——, “Sampling methods for the nyström method,” *Journal of Machine Learning Research*, vol. 98888, pp. 981–1006, June 2012.

Various column sampling methods for the Nyström kernel matrix approximation method are presented and discussed. Numerous experimentations are presented and different sampling methods are compared. The authors also present weight selection methods as an extension to the Ensemble Nyström method.

[8] M.-P. Lefranc, V. Giudicelli, Q. Kaas, E. Duprat, J. Jabado-Michaloud, D. Scaviner, C. Ginestoux, O. Clément, D. Chaume, and G. Lefranc, “Imgt, the international immunogenetics information system®,” *Nucleic Acids Research*, vol. 33, no. suppl 1, pp. D593–D597, 2005.

A database containing MHC molecule and T-Cell sequences. This is useful when extracting the pseudo-sequences of the MHC alleles to train a pan-specific predictor.

[9] M. Nielsen, S. Justesen, O. Lund, C. Lundegaard, and S. r. Buus, “NetMHCIIpan-2.0 - Improved pan-specific HLA-DR predictions using a novel concurrent alignment and weight optimization training procedure.” *Immunome research*, vol. 6, no. 1, p. 9, Jan. 2010.

A state-of-the-art pan-specific MHC-II-peptide binding affinity prediction method based on artificial neural networks. The authors show experimental results on many benchmark datasets and provide the dataset that was used to train their method. This training dataset, obtained the Immune Epitope Database, consists of 33931 measured binding affinities for 24 different MHC-II alleles. Since we are interested in using kernel methods for the richness of the feature space generated by kernels, the size of the dataset introduces the need for large scale kernel methods. Moreover, cross-loci pan-specific models are discussed. The authors put forward that for MHC-II, this task is complicated by the fact that only two allelic versions of HLA-DRA exist. Nevertheless, they mention that for HLA-DP and HLA-DQ, this task will be achievable in a near future, when more peptide binding data becomes available.

[10] M. Nielsen, C. Lundegaard, T. Blicher, B. Peters, A. Sette, S. Justesen, S. r. Buus, and O. Lund, “Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence : NetMHCIIpan.” *PLoS computational biology*, vol. 4, no. 7, p. e1000107, Jan. 2008.

The ancestor of the NetMHCIIpan-2.0 method. This is one of the first pan-specific models for MHC-II-peptide binding affinity prediction. Guidelines for the extraction of HLA-II pseudo-sequences are presented. Such pseudo-sequences are required to train pan-specific predictors of HLA-II, therefore the knowledge contained in this article is of great interest for someone intereted in training such a predictor.

[11] A. Patronov and I. Doytchinova, “T-cell epitope vaccine design by immunoinformatics,” *Open Biology*, vol. 3, no. 1, 2013.

A thorough review of the different types of computational methods available for vaccine development. A well structured introduction presents the different type of vaccines and a new branch of computer science and biology called *Immunoinformatics*.

[12] B. Peters, J. Sidney, P. Bourne, H.-H. Bui, S. Buus, G. Doh, W. Fleri, M. Kronenberg, R. Kubo, O. Lund, *et al.*, “The immune epitope database and analysis resource : from vision to blueprint,” *PLoS biology*, vol. 3, no. 3, p. e91, 2005.

This article presents the Immune Epitope database (IEDB) and the related analysis tools. The IEDB is currently one of the most well-maintained source of MHC-peptide epitope data. It contains over 200000 quantitative measurements of the binding affinity of peptides and MHC molecules for humans. For all the MHC-peptide complexes, the sequences of the MHC molecules and of the peptides are provided, which makes it a valuable source of data to train machine learning predictors at predicting the binding affinity of MHC molecules and peptides.

[13] C. Saunders, A. Gammerman, and V. Vovk, "Ridge regression learning algorithm in dual variables," in *(ICML-1998) Proceedings of the 15th International Conference on Machine Learning.* Morgan Kaufmann, 1998, pp. 515–521.

The Kernel Ridge Regression learning algorithm. This is currently a state-of-the-art algorithm for regression. It is a kernelized version of the well-known Ridge Regression algorithm. We will use this algorithm to train a pan-specific MHC-II-peptide binding affinity predictor.

[14] N. C. Toussaint and O. Kohlbacher, "Towards in silico design of epitope-based vaccines," *Expert Opinion on Drug Discovery*, vol. 4, no. 10, pp. 1047–1060, 2009.

This article is an extensive review of the computational methods used for the design of epitope based vaccines. It provides a clear description of the epitope-based vaccine design process and the challenges associated to each of its steps, which are : antigen identifcation, epitope discovery, epitope selection and vaccine assembly. The authors particularly insist on the epitope discovery step by reviewing the most successful computational methods used to predict antigen epitopes. Finally, the open problems of this field, including T Cell recognition of peptide-MHC complexes, are explained and discussed.

[15] C. Widmer, N. Toussaint, Y. Altun, O. Kohlbacher, and G. Rätsch, "Novel machine learning methods for mhc class i binding prediction," *Pattern Recognition in Bioinformatics*, pp. 98–109, 2010.

Introduces the use of physicochemical properties of amino acids in string kernels for protein sequences.

[16] C. K. Williams and M. Seeger, "Using the nystrom method to speed up kernel machines," *Advances in neural information processing systems*, pp. 682–688, 2001.

The Nyström method for kernel matrix approximation. This is a state-of-the-art method for approximating a kernel matrix based on its eigenfunctions. It allows to reduce the complexity of Kernel Ridge Regression from $O(n^3)$ to $O(m^2p)$.

[17] L. Zhang, K. Udaka, H. Mamitsuka, and S. Zhu, "Toward more accurate pan-specific MHC-peptide binding prediction : a review of current methods and tools." *Briefings in bioinformatics*, Sept. 2011.

An extensive review of the current state-of-the-art pan-specific binding affinity predictors for peptides and MHC molecules. This article explains the benefits of using pan-specific models, describes multiple different methods to train such models and presents benchmark datasets.

Interesting

[18] "MHC class II epitope predictive algorithms." *Immunology*, vol. 130, no. 3, pp. 319–28, July 2010.

The quality of the Immune Epitope Database is discussed. The authors raise problems on the data acquisition process for MHC-II. They also attempt to elucidate why *in silico* prediction methods are accurate for MHC-I, but tend to stabilize at a low level of accuracy for MHC-II.

[19] A. J. Bordner and H. D. Mittelmann, “Multirta : A simple yet reliable method for predicting peptide binding affinities for multiple class ii mhc allotypes,” *BMC bioinformatics*, vol. 11, no. 1, p. 482, 2010.

A pan-specific method for the prediction of MHC-II-peptide binding affinity prediction.

[20] ——, “Prediction of the binding affinities of peptides to class ii mhc using a regularized thermodynamic model,” *BMC bioinformatics*, vol. 11, no. 1, p. 41, 2010.

A single target binding affinity prediction method for MHC-II and peptides.

[21] T. Hastie, R. Tibshirani, and J. H. Friedman, *The elements of statistical learning : data mining, inference, and prediction.* New York : Springer-Verlag, 2001.

A classical book on statistical machine learning.

[22] X. Hu, W. Zhou, K. Udaka, H. Mamitsuka, and S. Zhu, “MetaMHC : a meta approach to predict peptides binding to MHC molecules.” *Nucleic acids research*, vol. 38, no. Web Server issue, pp. W474–9, July 2010.

An ensemble method of MHC-I and MHC-II binding affinity predictors. The authors show that their aggregated predictor outperforms individual predictors.

[23] L. Jacob and J.-P. Vert, “Efficient peptide–mhc-i binding prediction for alleles with few known binders,” *Bioinformatics*, vol. 24, no. 3, pp. 358–366, 2008.

The first kernel based pan-specific approach to MHC peptide binding affinity prediction. The authors present the feature space tensor product to combine multiple kernels.

[24] E. M. Lafuente and P. A. Reche, “Prediction of mhc-peptide binding : a systematic and comprehensive overview,” *Current pharmaceutical design*, vol. 15, no. 28, pp. 3209–3220, 2009.

Another review of MHC-peptide binding affinity prediction and a discussion of open problems in this field.

[25] P. Machart, T. Peel, L. Ralaivola, S. Anthoine, and H. Glotin, “Stochastic low-rank kernel learning for regression,” *arXiv preprint arXiv :1201.2416*, 2012.

An ensemble based low rank approximation method for regression.

[26] A. Patronov and I. Doytchinova, “T-cell epitope vaccine design by immunoinformatics,” *Open Biology*, vol. 3, no. 1, 2013.

A thorough review of the different types of computational methods available for vaccine development. A well structured introduction presents the different type of vaccines and a new branch of computer science and biology called *Immunoinformatics.*

[27] H.-G. Rammensee, J. Bachmann, N. P. N. Emmerich, O. A. Bachor, and S. Stevanović, “Syfpeithi : database for mhc ligands and peptide motifs,” *Immunogenetics*, vol. 50, no. 3, pp. 213–219, 1999.

A database of MHC ligands.

[28] C. E. Rasmussen and C. K. Williams, *Gaussian processes for machine learning.* MIT press Cambridge, MA, 2006, vol. 1.

Multiple methods for handling large datasets are described.

[29] A. Sette and R. Rappuoli, “Reverse vaccinology : developing vaccines in the era of genomics,” *Immunity*, vol. 33, no. 4, pp. 530–541, 2010.

The importance of predicting T-Cell epitopes is discussed and a figure summarizing the vaccine design pathway is provided.

[30] C. Widmer, N. Toussaint, Y. Altun, O. Kohlbacher, and G. Rätsch, “Novel machine learning methods for mhc class i binding prediction,” *Pattern Recognition in Bioinformatics*, pp. 98–109, 2010.

Introduces the use of physicochemical properties of amino acids in string kernels for protein sequences.

[31] G. L. Zhang, P. Bradley, N. Jojic, Y. Kim, O. Kohlbacher, C. Lundegaard, C. A. Magaret, M. Nielsen, C. Yanover, S. Zhu, *et al.*, “Machine learning competition in immunology–prediction of hla class i molecules,” *Journal of immunological methods*, vol. 30, no. 374, pp. 1–2, 2011.

[32] G. L. Zhang, H. H. Lin, D. B. Keskin, E. L. Reinherz, and V. Brusic, “Dana-farber repository for machine learning in immunology,” *Journal of immunological methods*, vol. 374, no. 1, pp. 18–25, 2011.

The Dana-Farber repository for machine learning was developed to bridge the gap between computational biologists and the machine learning community. To promote the collaboration between these two communities, they periodically organize machine learning competitions based on experimental biological data. The most recent competition was the Machine Learning in Computational Biology 2012, which was won by Giguère et al. from Laval University.

[33] H. Zhang, Z. Wang, and L. Cao, “Fast nyström for low rank matrix approximation,” in *Advanced Data Mining and Applications*, ser. Lecture Notes in Computer Science, S. Zhou, S. Zhang, and G. Karypis, Eds. Springer Berlin Heidelberg, 2012, vol. 7713, pp. 456–464.

An alternative to the Nyström method that does not require to compute the SVD decomposition of the square matrix W. Instead, it uses randomized principal component analysis, which is less computationally demanding.

[34] L. Zhang, Y. Chen, H.-S. Wong, S. Zhou, H. Mamitsuka, and S. Zhu, “TEPITOPEpan : extending TEPITOPE for peptide binding prediction covering over 700 HLA-DR molecules.” *PloS one*, vol. 7, no. 2, p. e30483, Jan. 2012.

A pan-specific matrix based MHC-II-peptide binding affinity prediction method.

[35] Q. Zhang, P. Wang, Y. Kim, P. Haste-andersen, J. Beaver, P. E. Bourne, H.-h. Bui, S. Buus, S. Frankild, J. Greenbaum, O. Lund, C. Lundegaard, M. Nielsen, J. Ponomarenko, A. Sette, Z. Zhu, and B. Peters, “Immune epitope database analysis resource,” vol. 36, no. May, pp. 513–518, 2008.

A set of tools for the prediction of T-Cell and B-Cell epitopes provided by the Immune Epitope Database.

Other

[36] “Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices.” *Nature biotechnology*, vol. 17, no. 6, pp. 555–61, June 1999.

[37] L. a. Abriata, M. L. M. Salverda, and P. E. Tomatis, “Sequence-function-stability relationships in proteins from datasets of functionally annotated variants : the case of TEM β-lactamases.” *FEBS letters*, vol. 586, no. 19, pp. 3330–5, Sept. 2012.

[38] R. Albert, “Scale-free networks in cell biology.” *Journal of cell science*, vol. 118, no. Pt 21, pp. 4947–57, Nov. 2005.

[39] M. Ayoub and D. Scheidegger, “Peptide drugs , overcoming the challenges , a growing business,” *Chemistry Today*, vol. 24, no. 4, pp. 46–48, 2006.

[40] A. Bairoch, R. Apweiler, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O’Donovan, N. Redaschi, and L.-S. L. Yeh, “The Universal Protein Resource (UniProt) ,” *Nucleic Acids Research*, vol. 33, no. suppl 1, pp. D154–D159, Jan. 2005.

[41] A. J. Bordner, “Towards universal structure-based prediction of class ii mhc epitopes for diverse allotypes,” *PloS one*, vol. 5, no. 12, p. e14383, 2010.

[42] A. P. Bradley, “The use of the area under the ROC curve in the evaluation of machine learning algorithms,” *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[43] C. J. Bryson, T. D. Jones, and M. P. Baker, “Prediction of immunogenicity of therapeutic proteins : validity of computational tools,” *BioDrugs*, vol. 24, no. 1, pp. 1–8, 2010.

[44] C. Lundegaard, O. Lund, C. Keşmir, S. Brunak, and M. Nielsen, “Modeling the adaptive immune system : predictions and simulations,” *Bioinformatics*, vol. 23, no. 24, pp. 3265–3275, 2007.

[45] M. Nielsen and O. Lund, “NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction.” *BMC bioinformatics*, vol. 10, p. 296, Jan. 2009.

[46] P. Oyarzún, J. J. Ellis, M. Bodén, B. Kobe, *et al.*, “Predivac : Cd4+ t-cell epitope prediction for vaccine design that covers 95% of hla class ii dr protein diversity,” *BMC bioinformatics*, vol. 14, no. 1, p. 52, 2013.

[47] T. Pahikkala, H. Suominen, and J. Boberg, “Efficient cross-validation for kernelized least-squares regression with sparse basis expansions,” *Machine Learning*, vol. 87, no. 3, pp. 381–407, June 2012.

[48] A. W. Purcell, J. McCluskey, and J. Rossjohn, “More than one reason to rethink the use of peptides in vaccine design.” *Nature reviews. Drug discovery*, vol. 6, no. 5, pp. 404–14, May 2007.

[49] J. Sidney, a. Steen, C. Moore, S. Ngo, J. Chung, B. Peters, and a. Sette, “Five HLA-DP Molecules Frequently Expressed in the Worldwide Human Population Share a Common HLA Supertypic Binding Specificity,” *The Journal of Immunology*, vol. 184, no. 5, pp. 2492–2503, Feb. 2010.

[50] P. L. Toogood, “Inhibition of protein-protein association by small molecules : approaches and progress.” *Journal of medicinal chemistry*, vol. 45, no. 8, pp. 1543–58, May 2002.

[51] J. a. Wells and C. L. McClendon, “Reaching for high-hanging fruit in drug discovery at protein-protein interfaces.” *Nature*, vol. 450, no. 7172, pp. 1001–9, Dec. 2007.

[52] C. Widmer, M. Kloft, N. Görnitz, and G. Rätsch, “Efficient Training of Graph-Regularized Multitask SVMs,” in *Proceedings of ECML 2012*, ser. Lecture Notes in Computer Science. Springer, 2012.

[53] C. Widmer, N. Toussaint, Y. Altun, O. Kohlbacher, and G. Rätsch, “Novel machine learning methods for mhc class i binding prediction,” *Pattern Recognition in Bioinformatics*, pp. 98–109, 2010.

[54] H. Zhang, O. Lund, and M. Nielsen, “The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities : application to MHC-peptide binding.” *Bioinformatics (Oxford, England)*, vol. 25, no. 10, pp. 1293–9, May 2009.