Detection and characterization of active compounds based on Random Matrix Theory

Magdalena Wiercioch

Jagiellonian University, Faculty of Physics, Astronomy and Applied Computer Science - Department of Information Technologies, Lojasiewicza 11, 30-348 Cracow

Introduction

- Predicting the biological activities of new compounds is challenging.
- Ligand-based statistical approaches are not very successful because of noise caused by undersampling. It shows that the number of molecules known to be active or inactive is vastly less than the number of possible chemical features that might determine binding.
- Machine learning methods infer the optimal representation of molecules directly from data. However, it does not resolve this undersampling problem, as the available data are usually significantly less than the number of parameters in the model.

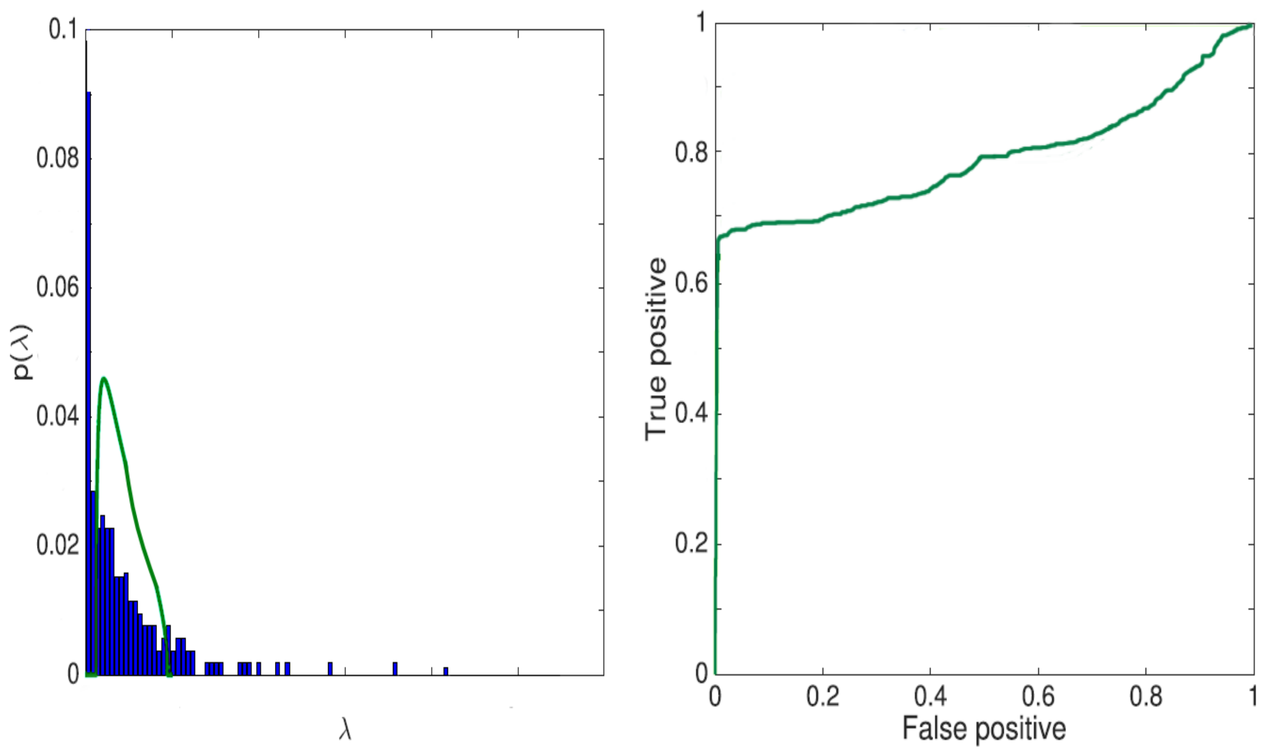
Methodology

- Molecular descriptors (fingerprints) are typically constructed by first representing a molecule as a 2D molecular graph and then considering all possible bond paths (contiguous atoms connected by chemical bonds) within the molecule. One can make an assumption that only identical molecules would share the same bond paths, and similar molecules share most bond paths.
- Lee et al. have shown that for a randomly chosen set of molecules, the eigenvalue distribution of the covariance matrix of chemical descriptors agrees with the canonical Marčenko-Pastur (MP) distribution [2] of RMT, expected in the absence of any significant signal [1].
- If one considers descriptors of pharmacologically similar molecules, i.e., those that bind to the same protein receptor, then part of the eigenvalue spectrum agrees with the MP distribution. Also, there are eigenvalues that deviate from it significantly. These eigenvalues, and their corresponding eigenvectors, describe the statistically significant signals.

Our contribution

- We present an extension of the work of Lee et al. [1] that is inspired by Random Matrix Theory.
- Classification of the molecule m provided by $\arg\min \mathrm{D(m, A)}$ and $\arg\max \mathrm{D(m, I)}$, where $D(m, Act) = ||m - \sum_{i=1}^{k}[\frac{act_i \cdot (m-\mu)}{\sigma}]act_i||_2$ and $D(m, Ina) = ||m - \sum_{i=1}^{k}[\frac{ina_i \cdot (m-\mu)}{\sigma}]ina_i||_2$.

Preliminary Results



Target	Our approach	Random matrix	SVM	Naive Bayes
Q14416	**0.77**	0.73	0.68	0.56
Q9HC97	**0.69**	0.66	0.64	0.53
Q99835	**0.76**	0.7	0.66	0.63
P50406	**0.81**	0.77	0.73	0.59
P51677	0.76	0.78	0.74	0.55
P21452	0.72	0.7	0.75	0.59

Conclusions

- Our model outperforms other state-of-the-art models.
- This is still work in progress.

References

[1] Michael P Brenner, Lucy J Colwell, et al. Predicting protein–ligand affinity with a random matrix framework. *Proceedings of the National Academy of Sciences*, 113(48):13564–13569, 2016.

[2] V.A. Marčenko and Leonid Pastur. Distribution of eigenvalues for some sets of random matrices. *Math USSR Sb*, 1:457–483, 01 1967.

Contact Information

magdalena.wiercioch@uj.edu.pl