A new large training database and Komet,an efficient algorithm for Targets Identification after phenotypic drug screening

Gwenn Guichaoua[1,2,3], Chloé Azencott [1,2,3], Véronique Stoven[1,2,3]

[1] Institut Curie, PSL Research University, 75428 Paris, France
[2] Center for Computational Biology, Mines Paris, PSL Research University, 75006 Paris, France
[3] INSERM U900, 75005 Paris, France

MINES PARIS | PSL
CBIO – Centre for Computational Biology

Context

- Find new therapeutic strategies using phenotypic (tumor cells survival) screens of molecules.
- Phenotypic screens provide hit molecules but not their targeted proteins/ mechanism of action. These are necessary for rational drug design.
- Testing the entire human proteome against hit molecules *in vitro* is impossible -> need for *in silico* methods to predict protein-ligand interactions.
- Motivation: analysis of a phenotypic survival screen of particular Breast Cancer tumor cells for which the hit molecules are 20 kinase inhibitors, which are known to be non-specific.

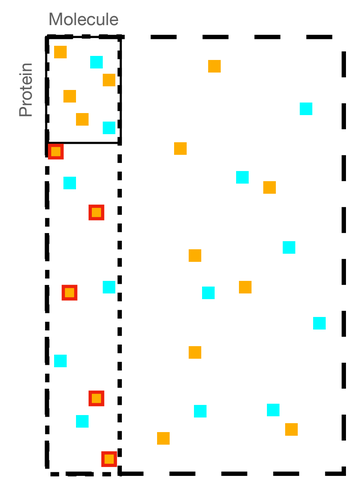
Goal: Predict the proteins targeted by the 20 hits and which may be responsible for the phenotype

Prediction of protein-ligand interactions

Supervised learning Binary classification problem

Input: database of interactions 1 -1

Output: predicted interactions



Challenges

The most complete training base
The largest, the most consensual
With direct interactions and negative interactions

The most efficient algorithm
In all scenarii of prediction, in a timely manner,
With reasonable computing resources

LCIdb, a new training base

Bioactivity database, extracted from 5 bioactivity databases: A Consensus Compound/Bioactivity Dataset for Data-Driven Design and Chemogenomics [Isigkeit *et al*, 2022]

	ChEMBL ID	PubChem ID	IUPHAR ID	Target	Activity type	Unit	Mean_C (0)	Mean_PC (9)	Activity check annotation	Ligand names	Structure check (Tanimoto)	Source
0	CHEMBL1448722	776051.0	NaN	alox15	pEC50	neg. log	4.9 *(1)	5.0 *(8)		2-(5-(3-bromophenyl)-1h-pyrazol-3-yl)-4-methyl...	match	chembl, pc
1	CHEMBL1279	77992.0	7191.0	htr1a	pKi	neg. log	7.2 *(1)	7.2 *(41)		(r)-6-methylamino-6,7,8,9-tetrahydro-5h-carbaz...	match	chembl, iuphar, pc, pd
2	CHEMBL146264	9830880.0	NaN	itgav	pIC50	neg. log	12.0 *(1)	8.6 *(1)	check activity data	(s)-3-{[5-(3-guanidino-phenyl)-thiophene-2-car...	match	chembl, iuphar, pc, pd
3	CHEMBL1279	77992.0	7191.0	htr1d	pKi	neg. log	8.4 *(1)	8.4 *(41)		(r)-6-methylamino-6,7,8,9-tetrahydro-5h-carbaz...	match	chembl, iuphar, pc, pd
4	CHEMBL1456115	80533.0	NaN	p2ry12	pIC50	neg. log	2.7 *(1)	2.7 *(55)		4-nitrobenzotriazole\nnsc-44657\n4-nitro-1h-be...	match	chembl, pc
5	CHEMBL1270660	80825.0	NaN	phlpp2	pIC50	neg. log	4.1 *(1)	4.2 *(1)		2-hydroxy-3-methyl-5-((4-((4-sulfophenyl)diaze...	no match (0.858)	chembl, pc

Preprocessing : For a (molecule,protein) pair

1. Remove pairs with multiple inconsistent bioactivities (difference > 1 log unit)
2. Remove molecule with different SMILES in different sources
3. Remove pairs for which non of IC50, Ki, Kd is available
4. Binarize interactions: measure = first Kd, then Ki, then IC50
   measure < 10 nM (10^{-7} M): interactions +
   measure > 100 microM (10^{-4} M): interactions -

Construction of a large new molecule/protein interactions dataset

2 069 proteins
274.515molecules
402.000 interactions +
50.000 interactions -

Komet, fast and efficient chemogenomic algorithm

Database for training a set of proteins $(p_\ell)_\ell$; a set of molecules $(m_k)_k$; a set of N positive/no interactions $I = I^+ \cup I^- = (\ell_i, k_i)_{i=1\dots N}$

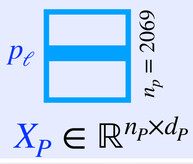
Choice of kernels [Scholkopf&al, 2004] *Morgan Fingerprint Kernel* $\kappa_M(m, m')$*:* similarity between molecules, *Local Alignment Kernel* $\kappa_P(p, p')$*:* similarity between proteins
Kernel κ *:* similarity between two pairs (m, p) and (m', p') defined by a Kronecker product $\kappa((m, p), (m', p')) = \kappa_M(m, m') \times \kappa_P(p, p')$

Problem Kronecker kernel K for training is too big for both storage and computation, and `sklearn` impraticable -> **From Kernels back to features**

Protein features

Singular value decomposition (SVD) of empirical kernel K_P

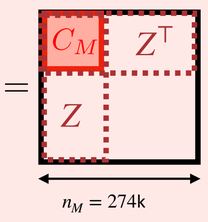
$K_P = U\text{diag}(\lambda)U^T = X_P X_P^T \qquad X_P = U\text{diag}(\sqrt{\lambda})$



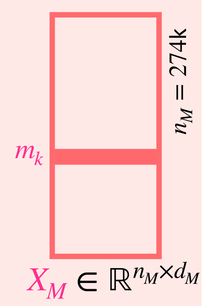
$X_P \in \mathbb{R}^{n_P \times d_P}$

Molecular features using Nystrom approximation *[Scholkopf et al, 1999]*

$K_M = \begin{pmatrix} C_M & Z^\top \\ Z & \end{pmatrix} \approx Z \times C_M^{-1} \times Z^\top$



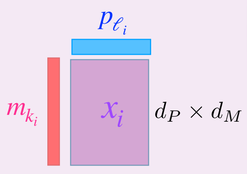
(SVD) $C_M = V\text{diag}(\mu)V^T \qquad X_M = ZV\text{diag}(1/\sqrt{\mu})$

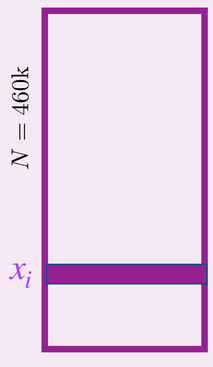


$X_M \in \mathbb{R}^{n_M \times d_M}$

Joint lifting with Kronecker kernel

$(p_{\ell_i}, m_{k_i}) \xrightarrow{\text{Tensor product}} x_i = m_{k_i} p_{\ell_i}^\top$

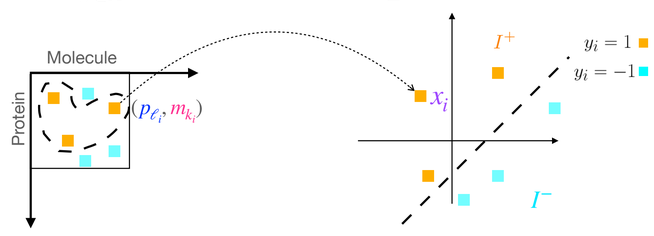




$X \in \mathbb{R}^{N \times (d_P \times d_M)}$

SVM classification in feature space

$$\min_{w \in \mathbb{R}^{d_P \times d_M}} L_{\text{Hinge}}(y, Xw) + \frac{\lambda}{2}\|w\|^2$$



Problem X is too big for both storage and computation of Xw

Komet : fast and efficient algorithm

Efficient computation [Airola&Pahikkala,2017]

2 Key ideas $(Xw)_i = \langle m_{k_i} p_{\ell_i}^\top, w \rangle = \langle m_{k_i}, \underbrace{W p_{\ell_i}}_{q_{j_k}} \rangle$

$(q_j)_{j=1}^{n_P}$ can be computed in only $n_P \times d_Z$ operations

	Explicit Xw	Implicit computation
Complexity	$N \times (d_P \times d_M)$	$n_P \times (d_P \times d_M) + N \times d_M$

Full batch BFGS method to solve the optimization problem

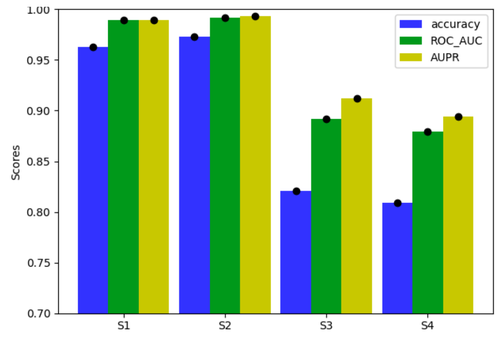
Code in PyTorch running on GPU https://komet.readthedocs.io

Results

Excellent performance

in different prediction scenarii [Playe *et al*, 2018]

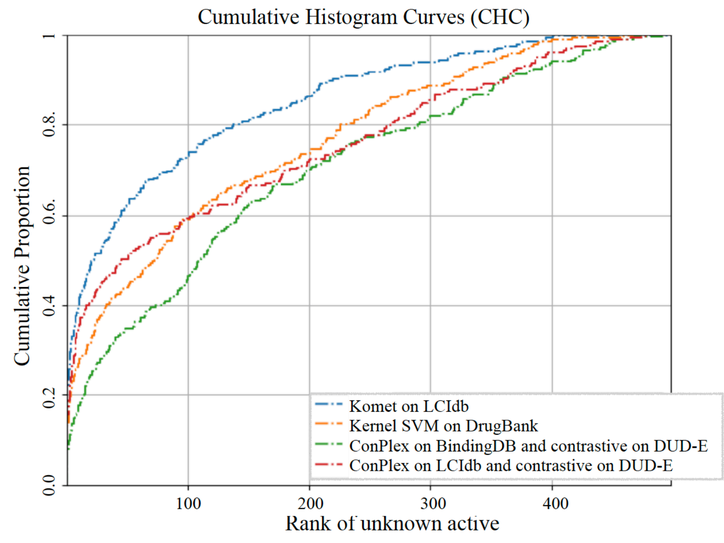
S1 Random
S2 Unseen Drugs in Test set
S3 Unseen Targets in Test set
S4 Unseen Drugs and Targets in Test set



Better AUPR to state-of-the-art Deep Learning algorithms [Singh *et al*, 2023]

Name	Drugs	Targets	Train/Val/Test	Fast Large Scale SVM	ConPLex [Singh, 2023]	MolTrans [Huang, 2021]
BIOSNAP	4,510	2,181	19,238/2,748/5,492	0.940 ± 0.001	0.921 ± 0.002	0.893 ± 0.001
Unseen_drugs			19,151/2,736/5,593	0.913 ± 0.002	0.899 ± 0.003	0.871± 0.002
Unseen_targets			19,375/2,768/5,340	0.891 ± 0.001	0.863 ± 0.031	0.683 ± 0.005
BindingDB	7,165	1,254	12,668/6,644 */13,289 *	0.667 ± 0.005	0.669 ± 0.003	0.611 ± 0.004
LCIdb	274,515	2,069	644,060/47,304/96,608	0.989 ± 0.0005 (15")	0.969± 0.002 (1630")	0.9719± 0.001 (69838")
Unseen_drugs			627,768/57,328/112,656	0.993 ± 0.0002 (15")	0.978± 0.003 (1734")	0.970 9± 0.002(68400")
Unseen_targets			618,734/59,822/121,644	0.912 ± 0.001 (15")	0.894 ± 0.031 (1329")	0.598 9± 0.007 (64800")
Orphan			236,530/22,503/45,006	0.894 ± 0.0004 (8")	0.846 ± 0.003 (888")	0.562 9± 0.013 (25200")

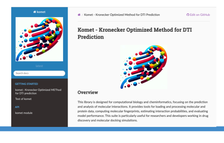
Recovering more out-of-scaffold hits than state-of-the-art Ligand-Based methods [Pinel *et al*, 2023]



Conclusion

We presented efficient tools to predict therapeutic targets/mechanisms of action after a phenotypic screen

- LCIdb: a new large molecule/protein interactions dataset to train ML algorithms
- Komet: Fast and state-of-the-art algorithm https://komet.readthedocs.io



References

1. Isigkeit *et al* (2022), Molecules
2. Scholkopf *et al* (2004), MIT press
3. Playe *et al* (2018), Plos One
4. Singh *et al* (2023), PNAS
5. Pinel *et al* (2023), Molecular Informatics
6. Airola, Pahikkala (2017), IEEE transactions on neural networks and learning systems

Project supported by the Île-de-France Region as part of the "DIM AI4IDF"