Applicability of the CGBVS Method in Predicting Ligands for Orphan Targets

Enzo Kawasaki[1,*], Chisato Kanai[1] and Atsushi Yoshimori[2]

[1]INTAGE Healthcare, Inc., NREG Midosuji Bldg., 3-5-7 Kawaramachi Chuo-ku, Osaka 541-0048 Japan.
[2]Institute for Theoretical Medicine, Inc., 26-1 Muraoka-Higashi 2-chome, Fujisawa, Kanagawa 251-0012, Japan.

Introduction

Drug-target interaction (DTI) analysis plays a crucial role in drug discovery. Recent progress in deep learning has led to advanced DTI models [1], enabling efficient screening of compounds and aiding in drug repositioning [2] by uncovering new therapeutic uses for existing medications.

Our group has continuously refined the CGBVS (Chemical Genomics-Based Virtual Screening) method for DTI prediction. Previously, we tested its accuracy on virtual orphan GPCR targets by omitting ligand data during training [3]. We employed the same approach in this study but we broaden the scope by also including kinases, ion channels and proteases.

Preparation and Validation of Virtual Orphan Predictive Models

In this study, we utilized chemogenomics-based virtual screening (CGBVS) [3] method to create virtual orphan models. The general protocol is illustrated in Fig 1.

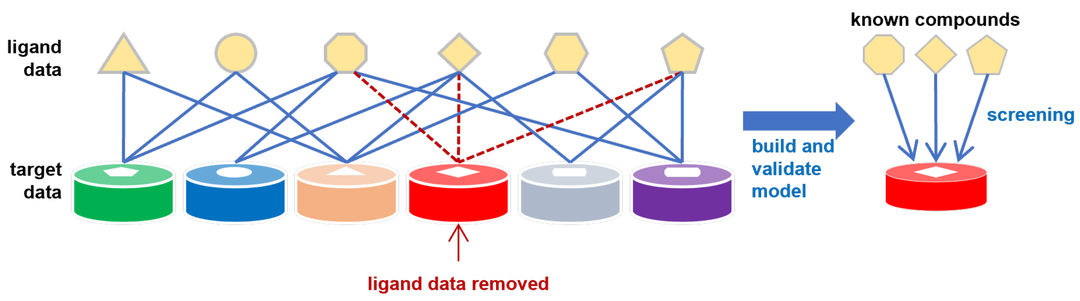

Figure 1. Process flowchart for this study. Red, dashed lines indicate removal of ligand data.

Targets of these models belong to 4 protein families (GPCR, kinases, ion channels and proteases) and for each protein family we selected 60 protein targets from whose training dataset we omitted ligand data. A total of 240 predictive models were created. Each model was used to screen a compound dataset. Results from the screening were used to calculate AUROC and applicability index. Information about the predictive models is shown in Table 1.

Table 1. Information about the predictive models used in this study

Target protein families:	GPCR, Kinase, Ion channel, Protease
Source DB:	ChEMBL rel. 33
Compound descriptors:	alvaDesc ver. 2
Protein descriptors:	Multiple Sequence Alignment (MSA)
Classifier:	Support vector machines (SVM)
Activity cutoff:	$\leq 10\mu$M

Applicability Index

We considered the applicability index $A(\boldsymbol{p}_i)$ of CGBVS to the virtual orphan target $\boldsymbol{p}_i$ to be proportional to the sum of the number of active ligands N_j of the neighboring proteins $\boldsymbol{p}_j$. So we defined it as

$$A\left(\boldsymbol{p}_i\right) = \sum_j w\left(K_{\mathrm{P}}\left(\boldsymbol{p}_i, \boldsymbol{p}_j\right)\right) N_j. \tag{1}$$

Here, w is a weight function whose argument is the value of the protein kernel function K_{P}, and its functional form is the sigmoid function

$$w(x) = \frac{1}{1 + \exp\left(-\alpha\left(x - r\right)\right)}. \tag{2}$$

The two parameters of the sigmoid function, α and r, are determined to maximize the Pearson's correlation coefficient between AUROC and $\log A$. The AUROC is calculated using the procedure described in a previous paper [3]. Bayesian optimization was used to perform the optimization of the correlation coefficient.

Relationship Between Reference Virtual Orphan Targets and Surrounding Proteins

Below are t-SNE (t-Distributed Stochastic Neighbor Embedding) representations of reference virtual orphan targets with their surrounding proteins. Similarity scores between a selected hit compound obtained via CGBVS and known ligands of the surrounding proteins are indicated in blue. The ligand data information of the surrounding proteins propagates to the orphan target enabling the prediction of potential ligands.

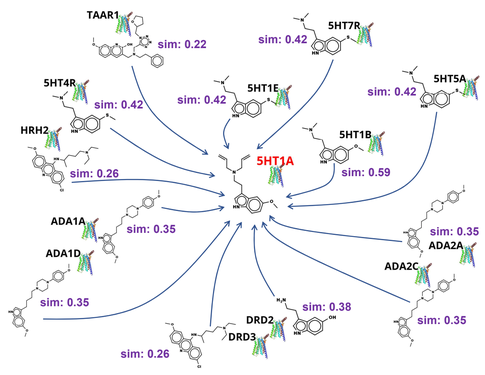

Figure 2. t-SNE representation of the GPCR 5HT1A and its surrounding proteins.

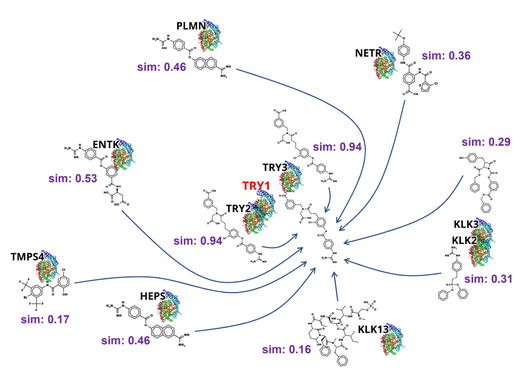

Figure 3. t-SNE representation of the protease TRY1 and its surrounding proteins.

Correlation of AUROC and logA

Scatter plots showing correlation of AUROC and logA of selected GPCRs, kinases, ion channels and proteases. GPCRs and proteases showed high correlation values followed by kinases with ion channels showing the least correlation.

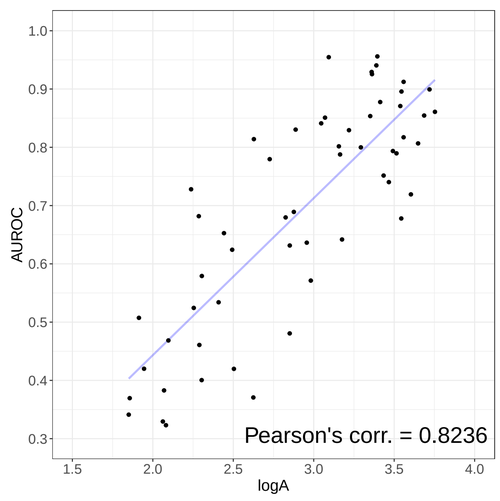

Figure 4. GPCR

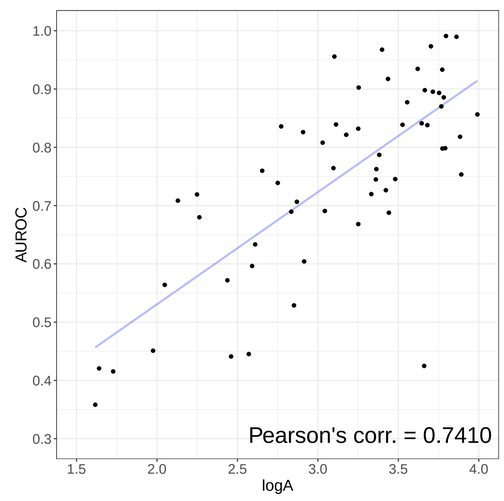

Figure 5. Kinase

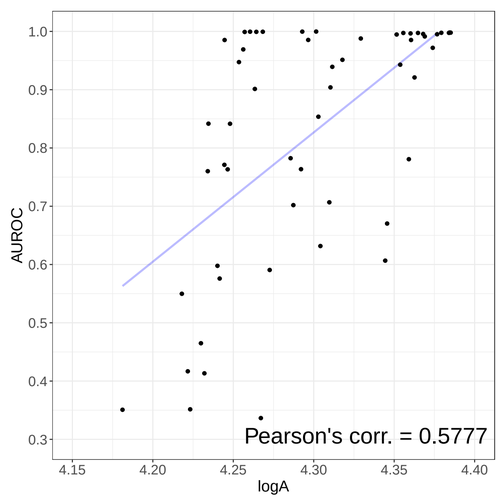

Figure 6. Ion channel

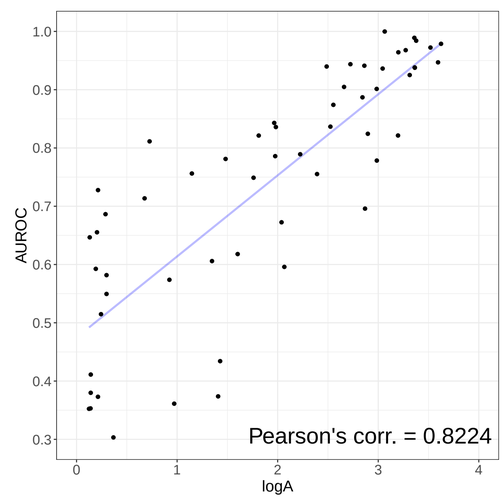

Figure 7. Protease

Orphan Targets with High Applicability Indices

We selected the proteins that have low number (0-10) of ligand data and exhibited high applicability indices. We believe that potential ligands of these orphan targets have a high probability of being predicted using the CGBVS method.

Table 2. List of orphan targets with possessing high applicability index

uid	ligand count	logA	accession	protein name	protein family
GLP2R_HUMAN	0	3.88	O95838	Glucagon-like peptide 2 receptor	GPCR
PTH2R_HUMAN	0	3.57	P49190	Parathyroid hormone 2 receptor	GPCR
SCTR_HUMAN	0	3.47	P47872	Secretin receptor	GPCR
ACTHR_HUMAN	0	3.33	Q01718	Adrenocorticotropic hormone receptor	GPCR
NPBW2_HUMAN	0	3.19	P48146	Neuropeptides BW receptor type 2	GPCR
FPR3_HUMAN	0	3.08	P25089	N-formyl peptide receptor 3	GPCR
VIPR1_HUMAN	1	3.36	P32241	Vasoactive intestinal polypeptide receptor 1	GPCR
CDK20_HUMAN	9	3.66	Q8IZL9	Cyclin-dependent kinase 20	Kinase
CDK10_HUMAN	10	3.72	Q15131	Cyclin-dependent kinase 10	Kinase
PLM_HUMAN	0	4.41	O00168	Phospholemman	Ionch channel
KCNE3_HUMAN	0	4.36	Q9Y6H6	Potassium voltage-gated channel subfamily E member 3	Ionch channel
KCNH6_HUMAN	0	4.35	Q9H252	Potassium voltage-gated channel subfamily H member 6	Ionch channel
KCNH7_HUMAN	0	4.34	Q9NS40	Potassium voltage-gated channel subfamily H member 7	Ionch channel
KCNE2_HUMAN	0	4.32	Q9Y6J6	Potassium voltage-gated channel subfamily E member 2	Ionch channel
TRY6_HUMAN	0	3.54	Q8NHM4	Putative trypsin-6	Protease
MMP27_HUMAN	0	3.32	Q9H306	Matrix metalloproteinase-27	Protease
CAN8_HUMAN	0	2.90	A6NHC0	Calpain-8	Protease
CAN11_HUMAN	0	2.82	Q9UMQ6	Calpain-11	Protease
CAN3_HUMAN	0	2.68	P20807	Calpain-3	Protease
NAPSA_HUMAN	0	2.51	O96009	Napsin-A	Protease
FOH1B_HUMAN	0	2.51	Q9HBA9	Putative N-acetylated-alpha-linked acidic dipeptidase	Protease
PRS29_HUMAN	0	2.48	A6NIE9	Putative serine protease 29	Protease
KLK15_HUMAN	0	2.45	Q9H2R5	Kallikrein-15	Protease
HTRA3_HUMAN	0	2.44	P83110	Serine protease HTRA3	Protease
MMP20_HUMAN	1	3.01 9	O60882	Matrix metalloproteinase-20	Protease
MMP24_HUMAN	1	2.73	Q9Y5R2	Matrix metalloproteinase-24	Protease
ECE2_HUMAN	3	2.50	P0DPD6	Endothelin-converting enzyme 2	Protease
TRYB2_HUMAN	8	2.84	P20231	Tryptase beta-2	Protease
CAN9_HUMAN	9	2.71	O14815	Calpain-9	Protease
KLK2_HUMAN	9	2.69	P20151	Kallikrein-2	Protease

Note: Due to space constraints, only a part of the ion channels and proteases are shown here.

Conclusions

- The existence of ligand information for the surrounding proteins influences the prediction of ligand for an orphan target.
- High correlation values between AUROC and logA for GPCR, kinase, and protease models suggest that the CGBVS technique is likely effective for predicting ligands of orphan targets in these categories. Conversely, the low correlation for ion channel models indicates that CGBVS may not be suitable for predicting ligands of orphan targets in this group.

References

[1] Heba Askr, Enas Elgeldawi, Heba Aboul Ella, Yaseen AMM Elshaier, Mamdouh M Gomaa, and Aboul Ella Hassanien. Deep learning in drug discovery: an integrative review and future challenges. *Artificial Intelligence Review*, 56(7):5975–6037, 2023.

[2] Lijun Cai, Jiaxin Chu, Junlin Xu, Yajie Meng, Changcheng Lu, Xianfang Tang, Guanfang Wang, Geng Tian, and Jialiang Yang. Machine learning for drug repositioning: Recent advances and challenges. *Current Research in Chemical Biology*, page 100042, 2023.

[3] Chisato Kanai, Enzo Kawasaki, Ryuta Murakami, Yusuke Morita, and Atsushi Yoshimori. Computational prediction of compound–protein interactions for orphan targets using cgbvs. *Molecules*, 26(17):5131, 2021.