Discovering Ligands for TRP Ion Channels Using Formal Concept Analysis

Mahito Sugiyama[1,2], Kentaro Imajo[1], Keisuke Otaki[1], and Akihiro Yamamoto[1]

[1] Graduate School of Informatics, Kyoto University
Yoshida Honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
{mahito, imajo, ootaki}@iip.ist.i.kyoto-u.ac.jp,
akihiro@i.kyoto-u.ac.jp

[2] Research Fellow of the Japan Society for the Promotion of Science

Abstract. In this paper, we propose an inductive approach to find candidates of ligands for *TRP ion channels* from databases, which play crucial roles for sensory transduction of living things and are actively studied in biology and biochemistry. To study properties of TRP channels biologically, *ligands* are key tools, which are chemical substances and activate or inhibit TRP channels by docking to them. However, finding a new ligand is difficult; choosing candidates of ligands relies on expert knowledge of biologists and test experiments *in vitro* and *in vivo* costs high in terms of time and money. Thus an *in silico* approach to find candidates of ligands helps biologists. Here we achieve this task by treating as *semi-supervised learning* from ligand databases and using SELF (SEmi-supervised Learning via FCA) recently proposed by two of the authors. SELF finds classification rules from mixed-type data including both discrete and continuous variables using FCA (Formal Concept Analysis). We show that SELF works well compared to other learning methods, and find candidates of ligands for TRP channels from more than thousand ligands stored in a database.

Keywords: Semi-supervised learning, Formal Concept Analysis, TRP ion channel, Ligand, Docking

1 Introduction

TRP (Transient Receptor Potential) *ion channels* form a class of ion channels, which are usually located on the plasma membrane, and they play a crucial role for sensory transduction. In particular, *ThermoTRPs*, a subset of TRP channels, are activated by changes in temperature [5]. Each channel has its own thermal thresholds and is considered as a "gate" of temperature sensation, such as *cold* or *hot* [2]. To experimentally analyze TRPs (in biological sense), biologists use *ligands*, which are chemical substances and activate (called agonist or activator) or inhibit (called antagonist or inhibitor) TRPs' response (Figure 1). Interestingly, each ligand has *selectivity*; *i.e.*, binding cites of ion channels to which it can dock is limited and this is why they are convenient for experiments. However, finding ligands is difficult. Choosing candidates of ligands relies on expert knowledge of biologists and experiments for testing ligands *in vitro* and *in vivo* costs high in terms of time and money. Thus an *in silico* approach to find candidates of ligands will help biologists.

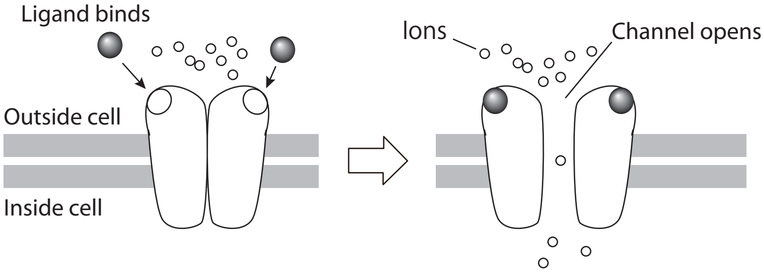

Fig. 1. Ligand-gated ion channels.

In this paper, we adopt an inductive, data mining, approach to find ligand candidates for TRPs from databases, and we use the framework of *semi-supervised learning* [3, 16] mainly studied in the machine learning community. Semi-supervised learning is a special form of classification, where a learning algorithm uses both labeled and unlabeled data to obtain a classification rule (a label is an identifier of a class). Commonly, only few labeled data are available since labeling data costs high in a real situation. Now only few ligands for TRPs are discovered, and lots of ligands for other receptors are available. Thus if we use ligands for TRPs and the other ligands as labeled and unlabeled data, respectively, semi-supervised learning fits to our goal.

Information about TRPs (and other ion channels) and ligands is donated to various databases, such as KEGG[3], and in this paper we use the IUPHAR database[4] [13]. In the database, for each ligand, we can know to which receptors it binds. Moreover, every ligand is characterized by seven attributes as follows: Hydrogen bond acceptors, Hydrogen bond donors, Rotatable bonds, Topological polar surface area, Molecular weight, XLogP, and No. Lipinski's rules broken. Here, forth, fifth, and sixth attributes are *real-valued* features, and the others are *nominal* features. Thus to learn classification rules for ligands from this database using the above seven attributes, a learning algorithm is required to handle *mixed-type data* including both discrete and continuous variables.

Now various semi-supervised learning methods are available, but most of them are for learning from data with real-valued features. Moreover, to the best of our knowledge, only the semi-supervised learning method SELF (SEmi-supervised Learning via Formal Concept Analysis) [14], proposed by two of the authors, can directly hadle mixed-type data. We therefore use SELF in this paper to obtain classification rules and discover licand candidates from ligand databases.

To date, no study treats mining of classification rules for ligands from databases. Some studies focus on predicting *affinity* of ligands, the strength of docking. Recently, the literature [1] proposed a machine learning approach to predict affinity, but we cannot know whether or not a ligand binds to a receptor. Most studies tried to construct a predictive model using domain-specific knowledge, such as the potential energy of a complex, the two-dimensional co-ordinates, and the free energy of binding [9]. However, to use such a method, some special background knowledge is required and results

[3] `http://www.genome.jp/kegg/`

[4] `http://www.iuphar-db.org/index.jsp`

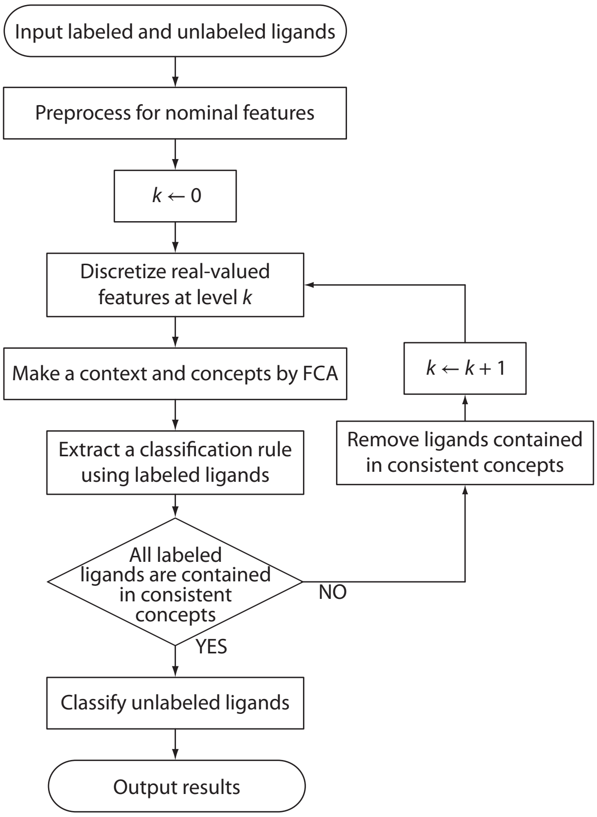
Input labeled and unlabeled ligands
Preprocess for nominal features
k ← 0
Discretize real-valued features at level k
Make a context and concepts by FCA
Extract a classification rule using labeled ligands
All labeled ligands are contained in consistent concepts
NO
YES
Remove ligands contained in consistent concepts
k ← k + 1
Classify unlabeled ligands
Output results

Fig. 2. A flowchart of classification by SELF. SELF learns classification rules from both labeled and unlabeled ligands (training data), and classify unlabeled ligands. We say that a concept is consistent if all labels contained in the concept are same.

depend on them. Our approach relies on only databases, hence the user do not need any background knowledge and can easily understand results.

This paper is organized as follows: Section 2 gives methods: an overview of FCA and SELF, and experimental settings. Section 3 describes results and discussion of experiments.

2 Methods

SELF algorithm. SELF [14] learns classification rules from ligand data using FCA. SELF allows missing values and labels in databases; this is why it can be viewed as a semi-supervised learning method. We briefly introduce FCA and SELF in the following.

FCA [4, 7] is a mathematical and algebraic method to derive a lattice structure, called a *concept lattice*, from a binary relation between objects and their attributes, called a *context* and given as a cross-table. In this study, each object corresponds to

Table 1. A subset of ligand database used for labeled data. Each ligand has seven attributes; Hydrogen bond acceptors (HBA), Hydrogen bond donors (HBD), Rotatable bonds (RB), Topological polar surface area (TPS), Molecular weight (MW), XLogP, and No. Lipinski's rules broken (NLR), and has a receptor to which it binds as a class label.

	HBA	HBD	RB	TPS	MW	XLogP	NLR	Receptor
allicin	1	0	5	61.58	162.02	0.24	0	TRPA1
allyl isothiocyanate	1	0	2	44.45	99.01	1.72	0	TRPA1
DOG	5	1	18	72.83	344.26	5.80	2	TRPC2
phosphatidylinositol	19	8	44	332.00	1022.49	9.87	4	TRPM4
menthol	1	1	1	20.23	156.15	3.21	0	TRPM8
eucalyptol	1	0	0	9.23	154.14	2.60	0	TRPM8
capsaicin	2	2	10	58.56	305.20	4.23	0	TRPV1
camphor	1	0	0	17.07	152.12	2.13	0	TRPV3
epoxyeicosatrienoic acid	3	1	14	49.83	320.24	6.58	2	TRPV4

a ligand, and SELF translates features of ligands into attributes of the context in the data preprocessing phase. Each concept obtained by FCA is a pair of objects and attributes with the *closed* property; *i.e.*, objects in a concept share a common subset of attributes and all attributes shared by the objects are in the concept. Many studies used FCA and the closed property for machine learning and knowledge discovery, such as classification [6] and association rule mining [10].

First SELF makes a context from a given mixed-type database using both labeled and unlabeled data, where we use level-wise discretization for continuous variables, and next it constructs the concept lattice from the context with FCA. Then SELF finds *maximal concepts* that are consistent with given class labels. Intuitively, their attributes correspond to the most general classification rules that explains a given labeled training data. We show a flowchart of SELF in Figure 2.

To efficiently find all concepts, we use the algorithm proposed by Makino and Uno [8], which is known to be one of the fastest algorithms. Their algorithm enumerates all maximal bipartite cliques in a bipartite graph that coincide with the concept. Its time complexity is $O(\varDelta^3)$, where $\varDelta$ denotes the maximum degree of a given bipartite graph. For empirical experiments, we use the program LCM[5] [15] to enumerate all concepts. As a result, time complexity of SELF is $O(nd)+O(\varDelta^3)+O(\varLambda)$, where n is the number of objects, d the number of attributes, and $\varLambda$ is the number of concepts at discretization level 1, since data preprocessing takes $O(nd)$, making concepts takes $O(\varDelta^3)$, and judging consistency of concepts takes less than $O(\varLambda)$.

Environment. All experiments were performed in R version 2.12.2 [11] since SELF was implemented in R. Note that LCM was implemented in C. We used Mac OS X version 10.6.5 with two 2.26-GHz Quad-Core Intel Xeon CPUs and 12 GB of memory.

[5] `http://research.nii.ac.jp/~uno/codes-j.htm`

Table 2. Results of accuracy (%). We used the all ligands as unlabeled data (SELF (ALL)), the subset of ligands which binds to TRPs (SELF (TRP)), or no unlabeled data (SELF). We used the decision tree-based method (Tree), SVM, and kNN ($k = 1, 5$).

SELF (ALL)	SELF (TRP)	SELF	Tree	SVM (RBF)	SVM (Pory)	1NN	5NN
0.52	0.48	0.37	0.18	0.39	0.43	0.50	0.34

Databases. We collected the entire 1,782 ligand data in the IUPHAR database[6] [13]. In the database, there are 44 ligands that binds to TRPs, where there exist seven TRPs: TRPA1, TRPC2, TRPM4, TRPM8, TRPV1, TRPV3, and TRPV4. From these ligands, we picked up nine ligands for labeled data shown in Table 1, which are known as famous and convenient ligands of TRPs for biological experiments. Other ligands for TRPs are used as a test data to evaluate performance of SELF. In the first experiment, we tested SELF in *transductive setting* [3], that is, we used both labeled and unlabeled data to obtain classification rules by SELF and predicted labels of unlabeled data. To measure the effectiveness of unlabeled ligand data, we performed three cases; use all ligands as unlabeled data, use the subset of ligands that binds to TRPs as unlabeled data, and use no unlabeled data. Moreover, to find new candidates of ligands for TRPs, we used all 44 ligands that binds to TRPs as labeled data in the second experiment.

Control Methods. As a control method for evaluation of SELF, we adopted the decision tree-based method implemented in R [12] since it can apply to mixed-type data. Note that this is supervised learning method and cannot use unlabeled data in the learning phase. Moreover, we applied SVM with the RBF and the polynomial kernels and k nearest neighbor method ($k = 1$ and 5) for reference by using only real-valued features.

3 Results and Discussion

Results are summarized in Table 2. These results show that unlabeled ligand data can be used effectively in the learning of classification rules. Moreover, if we use the all ligands for learning, SELF shows the best result compared to other learning methods, and the accuracy is more than 50 % despite there are seven classes. Notice that even though the nearest neighbor also records good results, we cannot obtain any classification rules. Our results are therefore valuable for finding new ligands by biological experiments.

In the second experiment, 79 classification rules were obtained by using the all ligands that bind to TRPs as labeled data and, by applying the rules, 762 candidates of ligands for TRPs were discovered from 1,782 ligands. These candidates are a novel result and can contribute to biological studies of TRP ion channels. Checking these candidates by actual biological experiments is a future work. Furthermore, this approach can be applied to any receptors, thereby discovering ligands for other receptors is an another interesting future work.

[6] `http://www.iuphar-db.org/index.jsp`

Acknowledgments

This work is inspired by insightful ideas of Professor Shigeo Kobayashi. This work was partly supported by Grant-in-Aid for Scientific Research (A) 22240010 and for JSPS Fellows 22·5714.

References

1. Ballester, P.J., Mitchell, J.B.O.: A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. Bioinformatics 26(9), 1169–1175 (2010)
2. Bautista, D.M., Siemens, J., Glazer, J.M., Tsuruda, P.R., Basbaum, A.I., Stucky, C.L., Jordt, S.E., Julius, D.: The menthol receptor TRPM8 is the principal detector of environmental cold. Nature 448(7150), 204–208 (2007)
3. Chapelle, O., Schölkopf, B., Zien, A. (eds.): Semi-Supervised Learning. MIT Press (2006), `http://www.kyb.tuebingen.mpg.de/ssl-book`
4. Davey, B.A., Priestley, H.A.: Introduction to lattices and order. Cambridge University Press, 2 edn. (2002)
5. Dhaka, A., Viswanath, V., Patapoutian, A.: TRP ion channels and temperature sensation. Annu. Rev. Neurosci. 29, 135–161 (2006)
6. Ganter, B., Kuznetsov, S.: Hypotheses and version spaces. In: de Moor, A., Lex, W., Ganter, B. (eds.) Conceptual Structures for Knowledge Creation and Communication. Lecture Notes in Computer Science, vol. 2746, pp. 83–95. Springer (2003)
7. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer (1998)
8. Makino, K., Uno, T.: New algorithms for enumerating all maximal cliques. Algorithm Theory-SWAT 2004 pp. 260–272 (2004)
9. Moitessier, N., Englebienne, P., Lee, D., Lawandi, J., Corbeil, C.R.: Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. British journal of pharmacology 153(S1), S7–S26 (2008)
10. Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L.: Efficient mining of association rules using closed itemset lattices. Information Systems 24(1), 25–46 (1999)
11. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing (2011), `http://www.R-project.org`
12. Ripley, B.D.: Pattern Recognition and Neural Networks. Cambridge University Press (1996)
13. Sharman, J.L., Mpamhanga, C.P., Spedding, M., Germain, P., Staels, B., Dacquet, C., Laudet, V., Harmar, A.J., NC-IUPHAR: IUPHAR-DB: New receptors and tools for easy searching and visualization of pharmacological data. Nucleic Acids Research 39, D534–D538 (2011), database Issue
14. Sugiyama, M., Yamamoto, A.: Semi-supervised learning for mixed-type data via formal concept analysis. In: Andrews, S., Polovina, S., Hill, R., Akhgar, B. (eds.) Conceptual Structures for Discovering Knowledge. Lecture Notes in Computer Science, vol. 6828, pp. 284–297 (2011)
15. Uno, T., Kiyomi, M., Arimura, H.: LCM ver. 3: Collaboration of array, bitmap and prefix tree for frequent itemset mining. In: Proceedings of the 1st International Workshop on Open Source Data Mining: Frequent Pattern Mining Implementations. pp. 77–86. ACM (2005)
16. Zhu, X., Goldberg, A.B.: Introduction to semi-supervised learning. Morgan and Claypool Publishers (2009)