Neurocomputing 206 (2016) 58–65

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Multi-fields model for predicting target–ligand interaction

Caihua Wang [a], Juan Liu [a,*], Fei Luo [a], Qian-Nan Hu [b,**]

[a] School of Computer, Wuhan University, Wuhan 430072, PR China
[b] Tianjin Institute of Industrial Biotechnology (TIB) of Chinese Academy of Sciences (CAS), Tianjin 300308, PR China

ARTICLE INFO

Article history:
Received 26 August 2015
Received in revised form
18 January 2016
Accepted 18 March 2016
Available online 30 May 2016

Keywords:
Protein representation
Ligand representation
Field interaction
Target–ligand interaction
Target–ligand network analysis

ABSTRACT

Predicting target–ligand interactions is a critical task of chemogenomics and plays a key role in virtual drug discovery. Moreover, it is important to take insights into the molecular recognition mechanisms between chemical substructures of ligands and binding sites of targets. In this work, we suppose the interaction between a ligand and a target is the result of the comprehensive effect of multiple fields between the ligand and the binding site of the target, and propose a multi-field interaction model (MFIM) to predict the target–ligand interaction. The evaluation result on the same data set shows that MFIM outperforms other two representative methods. The derived fragment interaction network is robust to the parameter fluctuation and the connections in the network are sparse. Moreover, the edge weights of the network might reflect the fragment interaction intensity and most of the significant edges are chemical interpretable.

1. Introduction

Target–ligand interactions play an initial role in cell signal conduction and predicting interactions between targets and ligands is the key task of chemogenomics. Through various high-throughput experimental projects for analyzing the genome, transcriptome and proteome, we are beginning to understand the genomic spaces populated by these protein classes. Meanwhile, the high-throughput screening of large-scale chemical compound libraries with various biological assays is enabling us to explore the chemical space [1,2]. However, our knowledge about the relationship between the chemical and genomic spaces is very limited [3,4]. Since experimental ways to determine target–ligand interactions are costly and time-consuming, there is a strong incentive to develop computational methods capable of detecting potential target–ligand interactions efficiently.

Traditional computational methods could be roughly classified into two categories: ligand-based ones and target-based ones [3]. In ligand-based approaches, a ligand is compared to the ligands known interacting with a given target to make the prediction. Therefore, these approaches require enough number of ligand binding with the given target. Molecular docking is a typical target-based approach, however, it is time-consuming and is tricky to find a score function to select the best conformation [5].

Both of the above types of methods only consider information on one side of the interaction pair whereas ignoring important information on the other side. In recent years, researchers have begun to take the viewpoint of target–ligand based chemogenomics approaches, taking information on both sides of the interaction pair into consideration. Bleakley and Yamanishi showed that the ensemble of ligand-based and target-based prediction models performed much better than only using single type of model [6]. Based on this fact, they proposed a bipartite local model (BLM). The BLM method has been further studied and improved by some researchers [7–9]. Leslie et al. first described the targets as full length sequences whose similarities were measured by the mismatch kernel [10,11]; and represented the ligands with PubChem fingerprints [12] whose similarities were measured by the Tanimoto kernel. Then they constructed the pairwise support vector machine (pSVM) to predict the target–ligand interactions. Vert et al. followed Leslie's framework except that they adopted the local alignment kernel [13] to compare target similarity. Jacob et al. also followed Leslie's framework, however they described targets (proteins) as EC numbers, and applied a hierarchy kernel (tree-based kernel) to compare target similarity [14]. Van Laarhoven et al. represented targets as sequences and ligands with molecular graphs. They applied Smith-Waterman score [15] to measure the target sequence similarity, and used SIMCOMP [4] to measure the similarity score between moleculars. They also got the target–ligand interaction network profiles and adopted Gaussian kernel to measure the network profile similarity. After that, they combined the Smith-Waterman score and Gaussian kernel to get the final target kernel, and combined the SIMCOMP score and Gaussian kernel to get the final ligand kernel. And then, they used the kernel ridge regression (RLS) [7] to make a prediction. Yamanishi et al. developed a supervised learning

* Corresponding author. Tel./fax: +86 027 68775711, +86 022 24828786.
** Corresponding author.
E-mail addresses: liujuan@whu.edu.cn (J. Liu), hu_qn@tib.cas.cn (Q.-N. Hu).
URLS: http://www.wangcaihua@whu.edu.cn (C. Wang), http://www.liujuan@whu.edu.cn (J. Liu), http://www.luofei@whu.edu.cn (F. Luo), http://www.hu_qn@tib.cas.cn (Q.-N. Hu).

http://dx.doi.org/10.1016/j.neucom.2016.03.079
0925-2312/

algorithm to infer unknown drug–target interactions by integrating the chemical space and genomic space into a unified space ("pharmacological space"), in which the higher probabilities a target and ligand interact with each other, the closer they are [4]. Besides, many researchers employ data-driven approaches to predict target–ligand interaction, as availability of many genomic and chemogenomic data sources. Wang et al. proposed a kernel-based method to predict drug–protein interactions by integrating multiple types of data [16–18]. Zhao and Li developed a computational framework, drugCIPHER, to infer drug–target interactions in a genome-wide scale [19].

Basic assumptions of most chemogenomic-based approach are twofold: compounds sharing some chemical similarity should also share targets and targets sharing similar ligands should share similar patterns/binding-sites [3]. Most of the previous work are based on the two assumptions [6–10,13,14]. Although those methods usually perform well on the data set, we know little about the interaction mechanisms, which means the algorithms work in a black box. In this work, inspired by QSAR [20] which decomposes the interactions into different fields, we propose a new model called multi-fields interaction model (MFIM) to give insight into the molecular recognition mechanisms between chemical substructures of ligands and binding sites of targets. Our hypotheses are (1) there are multiple types of fields around a molecular fragment. The interaction of two molecular fragments is the result of the comprehensive effect of multiple fields; (2) the fields around a binding-site/ligand originate in the superposition of the fields of their fragment. The target–ligand interaction is the result of the comprehensive effect of their fragment fields; (3) different types of fields are heterogeneous and heterogeneous fields cannot interact with each other. For instance, the electrostatic field does not interact with hydrophobic field. Based on the three hypotheses, we build our algorithm (Fig. 1). First, according to the target dictionary and the ligand dictionary, the binding sites of the targets and the ligands are encoded as binding site fragment vectors and ligand substructure vectors, forming the origin target and ligand spaces respectively. Then, we figure out the multiple field intensities around the molecular fragments $(\boldsymbol{U}, \boldsymbol{V})$, and the two origin spaces are mapped into target and ligand field spaces respectively. The target fields would interact with the homogeneous ligand fields and a pairwise field space is generated. Finally, we construct a classifier to predict site–ligand interactions in the pairwise field space. Moreover, we integrate the multiple fields of molecular fragments and obtain a fragment interaction network, which give insight into the molecular recognition mechanisms between chemical substructures of ligands and binding sites of targets.

2. Materials and methods

2.1. Data set and data representations

In this work, we use the data set constructed in our preliminary work [21]. There are totally 836 targets and 2710 corresponding ligands. Among those proteins, there are 782 targets with one binding site, which interact with 1988 ligands and form 2561 interaction pairs. The others possess multi-sites, which interact with 722 ligands and form 854 interaction pairs. The negative site–ligand pairs are generated as our preliminary work, totally 6830 target–ligand pairs are included in our data set [22,23].

Just as in our preliminary work [21], we represent each binding site as a 199-dimensional vector, in which each element denotes the occurring frequency of a set of trimers in the binding site (a trimer is a three-residue fragment of the binding site, and all possible trimers are clustered into 199 types according to their physical–chemical properties by Nagamine and Sakakibara [24,25]); we represent each ligand as a 413-dimensional binary vector, each element corresponds to the presence/absence of one chemical substructure (fragment) in the ligand dictionary (413 substructures in all).

2.2. Multi-field interaction model

Suppose there are p binding sites, and $\boldsymbol{s_i}$, an m-dimensional vector ($m=199$), denotes the $\boldsymbol{i}$-th binding site. Suppose there are q ligands, and $\boldsymbol{l_j}$, a n-dimensional vector ($n=413$), denotes the $\boldsymbol{j}$-th ligand. We use $\boldsymbol{D}=((i_1,j_1),(i_2,j_2),\ldots,(i_N,j_N))$ to denote all the site–ligand pairs in the data set, where N is the total number of site–ligand pairs ($N = 6830$),

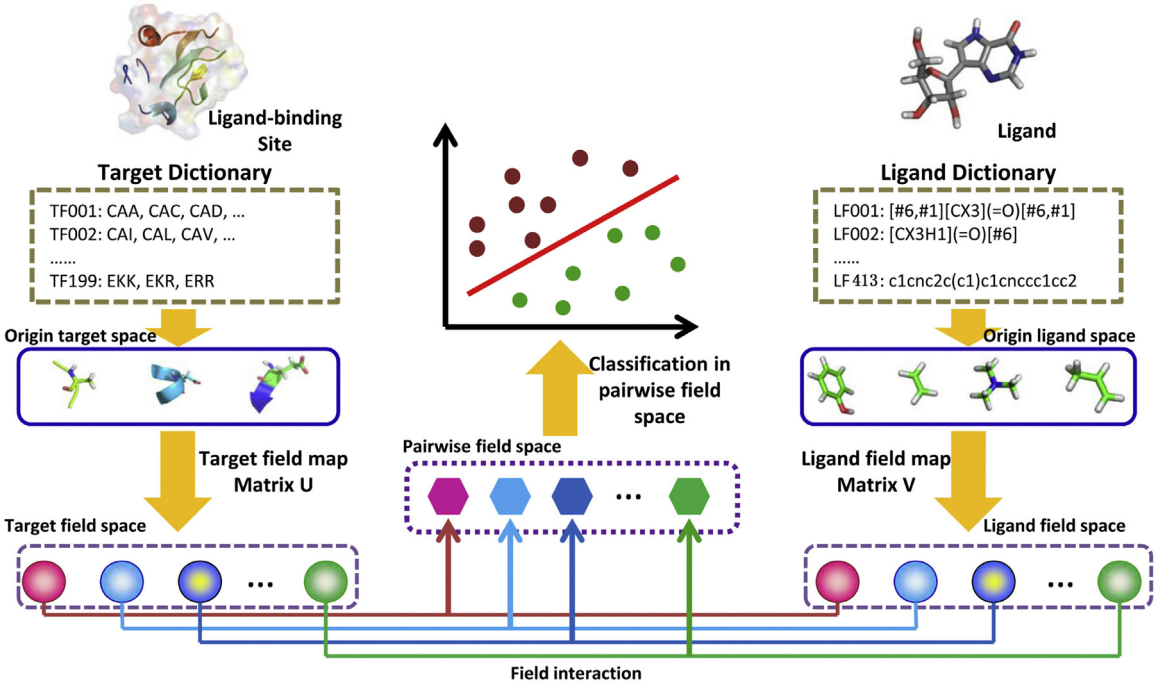

Fig. 1. Multi-field interaction model. First, based on the target dictionary and the ligand dictionary, the binding sites of the targets and the ligands are encoded as binding site fragment vectors and ligand substructure vectors, forming the target and ligand spaces respectively. Then, two spaces are mapped into target and ligand field spaces respectively. The target fields would interact with the homogeneous ligand fields and a pairwise field space is generated. Finally, we construct a classifier to predict site–ligand interactions in the pairwise field space.