Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

Statistical and machine learning approaches to predicting protein–ligand interactions

Lucy J Colwell

Data driven computational approaches to predicting protein–ligand binding are currently achieving unprecedented levels of accuracy on held-out test datasets. Up until now, however, this has not led to corresponding breakthroughs in our ability to design novel ligands for protein targets of interest. This review summarizes the current state of the art in this field, emphasizing the recent development of deep neural networks for predicting protein–ligand binding. We explain the major technical challenges that have caused difficulty with predicting novel ligands, including the problems of sampling noise and the challenge of using benchmark datasets that are sufficiently unbiased that they allow the model to extrapolate to new regimes.

Address
Department of Chemistry, Cambridge University, Cambridge, UK

Corresponding author: Colwell, Lucy J (ljc37@cam.ac.uk)

Current Opinion in Structural Biology 2018, **49**:123–128

This review comes from a themed issue on **Theory and simulation**

Edited by **Kresten Lindorff-Larsen** and **Robert Best**

https://doi.org/10.1016/j.sbi.2018.01.006

Introduction

Molecular recognition is a fundamental requirement of biological systems. The interactions between proteins and small molecules are central to biology, allowing cells to sense their surroundings and respond appropriately. Estimates place the number of small molecules that can be synthesized at $\approx 10^{60}$, yet just a small fraction of potential protein–ligand interactions have been explored. Finding novel interactions is of great importance to drug discovery and basic biology. Given the enormity of the search space, computational approaches can narrow down the possibilities. However, despite three decades of computational effort, biochemical experiments are still essential to determine the efficacy of ligand binding to a protein target [1,2]. The results of computational analysis have been decidedly mixed: it is challenging to use even experimentally well-characterized ligand–protein interactions to computationally design novel interactions [1,2], much less explore the vast space of possibilities.

There are three highly demanding tasks in protein–ligand binding prediction: *virtual screening* predicts whether a ligand binds to a given target; *affinity prediction* predicts the binding affinity; and *pose prediction* identifies the molecular interactions causing binding to occur. In this review we focus on the first; the others have been reviewed elsewhere [3,4]. Approaches to virtual screening can be categorized as physical or statistical. The idea of using first principles physical models to describe protein–ligand interactions is attractive, however timescale and computational resource constraints mean that simplified descriptions of features such as protein flexibility, and solvent are necessary. Docking algorithms are an important example of a coarse-grained physical model, however, even the most sophisticated versions cannot accurately reproduce large numbers of known interactions, much less predict new ones. The scoring functions used in such approaches can be empirical [5–9] or knowledge-based [10–14], and significant expertise is required to encode physico-chemical interactions through the use of hand-tuned features and parameters. Moreover, the results can be highly specific to the system that they are designed for [15].

Recently, the use of high throughput methods to screen large libraries of proteins and small molecules and quantify their interactions has made it possible to correlate activity with representations of proteins and small molecules, to infer predictive models. Techniques from machine learning and artificial intelligence have been introduced, allowing both the parameters and the model to be learned from the wealth of experimental data available in databases such as Chembl [16–19]. Increasingly publications are demonstrating that data driven approaches have the potential to make significant contributions to these problems [20••,21•,22•,23•,24–27].

Machine learning — potential and limitations

The aim of any machine learning or statistical approach is to identify patterns among training examples that can be used to make predictions outside the training set. An algorithm achieves this by mapping the training set representation to a space in which active and inactive ligands segregate — this mapping can be guided by physical models [13,23•] but is more often learned directly from the data without addition of extensive physico-chemical knowledge [21•,22•,26]. Once this mapping

has been learned, it is hoped that the location of new examples in this space — clustered with training set actives or inactives — will accurately predict their activity. To compare the performance of different algorithms, a test set of ligands with known activity is held-out during the training process. An algorithm that can make accurate predictions for this unseen data is presumed to extrapolate well.

In particular, approaches that use deep neural networks (DNNs) have been shown to make predictions on held-out test sets with eye-opening levels of accuracy, exceeding 0.95 AUC (Area under the Receiver Operator Characteristic) on benchmark datasets [22•,28]. However, the extent to which such results extrapolate is not yet clear — are they overfit to the training data [1,29,30••]? A number of studies posit that the test/train protocol is not as exacting as it might appear due to the non-uniformity of the distribution of ligands in chemical space. If training set actives are closer to test set actives than to training set inactives, by some metric that is not fully predictive of protein–ligand binding activity such as molecular weight, or number of ring systems, then the algorithm can appear to make accurate predictions for test set molecules without being able to extrapolate this predictive ability [31,19,32]. Machine learning performs best when abundant data drawn uniformly from the space of interest is available, but in this setting human chemists choose which molecules to work with, often based on clear similarities to known success stories [33,34].

Definitively showing that these approaches generalize is perhaps the outstanding challenge facing this field today. In the search for novel pharmaceuticals, the ability to predict the binding of ligands that are chemically distinct from those in the training data is highly valuable, but much more challenging for algorithms that are expert at identifying patterns among training set ligands. The goal of this article is to review statistical approaches to molecular recognition in the context of protein–ligand binding, focusing on recent results that exploit DNNs (see Figure 1). Here we briefly outline the basic steps of machine learning algorithm that predicts protein–ligand binding.

Molecular representation

There are almost as many choices for representation of the input data as there are for the machine learning algorithm employed [35,36]. The simplest involve counting the numbers of different heavy atoms present in a ligand, together with other features such as hydrogen bond donors/acceptors, chiral centres and ring systems [19,30••]. Some information about the chemical structure is retained by descriptors such as atom pairs or donor–acceptor pairs [37,38] where each element has the form (atom type *i*) — (distance in bonds) — (atom type *j*). More information is encoded by chemical fingerprints, for example MACCS keys [39] and ECFP fingerprints [40]; fixed length binary descriptors which can be generated by the package RDKit [41]. Here, each non-hydrogen atom is used as a centre from which fragments are generated by extending radially from the centre along bonds to neighbouring atoms; the maximum radius considered *N* is encoded in the name as ECFP2*N*. A unique identifier is assigned to each fragment, and the set of identifiers for molecules is mapped to a fixed length bit vector to yield the molecular fingerprint.

This abundance raises the question of which representation is most useful for different prediction tasks. Recently the suggestion has been made that it may be more effective to also learn the molecular representation itself, alongside the metric and corresponding embedding space used to distinguish active from inactive ligands [21•,22•,27,42]. However, counter-intuitively it has also been reported that the use of more complex molecular descriptors can result in little gain of predictive ability [30••,43].

Representation and sampling noise

One rational for the finding that more complex representations can result in little improvement is noise due to finite sampling. The basic premise of any predictive algorithm is that similarities among known interaction partners can reveal the requirements of the binding site, and thus predict novel interactions. A straightforward approach is to compile the set of ligands known to bind to a protein receptor of interest, and identify those features that show statistically significant enrichment among this set [44]. However, because there are only finitely many samples (i.e. known ligand binders), some features will be enriched purely by chance. This chance similarity increases with the number of variables, so representations that have more variables will lead to greater random similarity between features. For DNNs in particular, representations with thousands or even millions of features have recently been employed [24,21•,22•]; although these algorithms have the ability to share information between targets it is still important that the level of chance similarity between small molecules is quantified and accounted for.

This phenomena has been carefully studied in the field of random matrix theory, which provides a null distribution that describes the similarity between samples (ligands) that can be expected by chance due to finite sampling as a function of the number of samples available, and the number of variables present in the ligand descriptor [45,46•]. A simpler method for generating this null distribution involves computing the covariance matrices of multiple sets of *n* random ligands, where *n* is the sample size, using the same ligand descriptor for each set, to obtain the distribution of the largest entries that occur due to finite sampling noise. A similar approach can be taken for any measure of molecular similarity, defining a