論文の内容の要旨

DESCRIPTION OF MOLECULES AND SEARCH OF SPACE THAT THE DESCRIPTION DATA COMPOSE

（化合物の表現とその表現データが構成する空間の探索）

氏名　　襲　田　　勉

（本文）　Methods for molecule screening that sequentially search based on a statistical approach find drug candidates for a target protein effectively by converting the structure of chemical molecules with descriptors and prediction of molecular activity on a search space obtained by the conversion. Most of the molecules on the search space are inactive to a target protein. As a result, the search space is an imbalance data set that consists of a few active molecules data and a large number of inactive molecules data. Active molecules data do not always resemble to each other when the whole structures of active molecules do not resemble to each other. Moreover, the search of descriptors in addition to molecules data gives us more information about searching for drug candidates.

The accuracy of a ML algorithm on whole data sometimes increases as the accuracy on a majority class increases. Oppositely, the accuracy on a minority class decreases. In virtual screening, a minority class denotes active molecules and a majority class denotes the inactive molecules. The value of active molecules is much higher than that of inactive molecules. Therefore, the objective for improvement of

the accuracy on whole data set is sometimes inconsistent with the objectives for molecule screening. AL algorithms have the same problem because a ML algorithm builds hypotheses in the AL algorithms. Additionally, labeled data set sometimes include only negative data at the initial stage of virtual screening. In case of no positive data, ligand-based peptide mimetics design approach that finds out key structures for activity and converts the structure of peptides to the structure of small molecules based on the key structures sometimes gives the information about positive data. On an imbalanced data set, the error-correction based algorithm correct a hypothesis in favor of a class based on the error of query instances that is found out at the annotation of class.

Positive data that denote active molecules sometimes lose the similarity because the structures of active molecules are different to each other or because a descriptor sometimes converts two similar structures to two unlike features. In such case, AL algorithms do not always select instances on decision boundaries around unfound positive data. The substructure-pair descriptor describes the structure of peptides with having various kinds of scope to capture the structures that contribute to activity. The noise-tolerant AL algorithm sometimes finds out unfound decision boundaries based on the ratio of the density of labeled instance and the density of unlabeled instance around a query instance.

To find out important data and descriptors, the two-dimensional query strategy estimates the uncertainty of instances and a variety of each feature. Then, it selects an unlabeled instance according to a unified index between the uncertainty and the variety.

This thesis provides an approach to searching sequentially and statistically chemical space that includes the above three problems from the viewpoint of informatics and cheminformatics. As a result, the approach provides a method for molecule screening and enables the screening for various kinds of target proteins.