論 文 の 内 容 の 要 旨

論文題目　Molecule Identification from NMR Spectra with Machine Learning
（機械学習を用いたNMRスペクトルからの分子同定）

氏　　名　張　金哲

Introduction

Structure elucidation is the process of determining the chemical structure of a compound from a sample. It is a basic but essential task in biology and chemistry research. For organic compounds, structure elucidation is often involved with Nuclear magnetic resonance (NMR) spectroscopy. NMR spectroscopy is a technique that observes local magnetic fields around atomic nuclei. NMR spectroscopy machine excites the sample with a radio frequency pulse, a nuclear magnetic resonance response is then obtained. This nuclear magnetic resonance response, also known as a free induction delay (FID), can be converted into a spectrum with a Fourier Transformation.

On the NMR spectrum, functional groups in a molecule appear as peaks which characterizes themselves depending on their chemical environment in the molecule. Chemists can often readout the knowledge behind the chemical shifts where the peaks of some groups appear and assembly information together to obtain structural information of the sample molecule. Therefore, the approximate chemical structure of the target molecule can be confirmed from the NMR spectrum. In addition to structure elucidation by chemists, it is also possible to match observed NMR spectrum with existing NMR spectrum in the database. Although many previously observed NMR spectra are accumulated in public databases, they cover only a tiny fraction of the chemical space.

To overcome the limitation of current structure elucidation methods with NMR spectroscopy, an automated structure elucidation method independent to database content is desired.

Recent progress in machine learning has enabled the development of *de novo* molecule generators which are expected to design molecules with desired properties. Previously, our lab developed a molecule generator, ChemTS[1], which combines Monte Carlo tree search (MCTS) with a recurrent neural network (RNN), and successfully showed that ChemTS coupled with quantum chemical calculations can produce realistic molecules that have desired properties. So far, most *de novo* molecule generators have only been tested or applied on quantifiable chemical properties such as gaps between the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO). As ^{1}H NMR spectra are highly characteristic of individual compounds, we consider ^{1}H NMR spectra as one of its molecular properties.

In this thesis, we propose NMR-TS, a machine-learning-based python library, to automatically identify a molecule from its NMR spectrum. NMR-TS discovers candidate molecules whose NMR spectra match the target spectrum by using deep learning and density functional theory (DFT)-computed spectra. As a proof-of-concept, we identify prototypical metabolites from their computed spectra. After an average 5451 DFT runs for each spectrum, six of the nine molecules are identified correctly, and proximal molecules are obtained in the other cases. This encouraging result implies that *de novo* molecule generation can contribute to the fully automated identification of chemical structures.

Methods

The concept of this study is to generate a molecule which has a similar NMR spectrum to the target one of an unknown molecule, illustrated in Fig. 1. We use the simulate ^{1}H spectrum obtained by density functional theory (DFT) calculation and the number of Hydrogen and Carbon atoms as information of an unknown molecule (0: Target spectrum). The proposed method, called NMR-TS, searches a molecule which have the same NMR spectrum by using ChemTS as a *de novo* molecule generator and DFT calculations. ChemTS generates molecules based on the MCTS algorithm using the SMILES format of molecule representation.

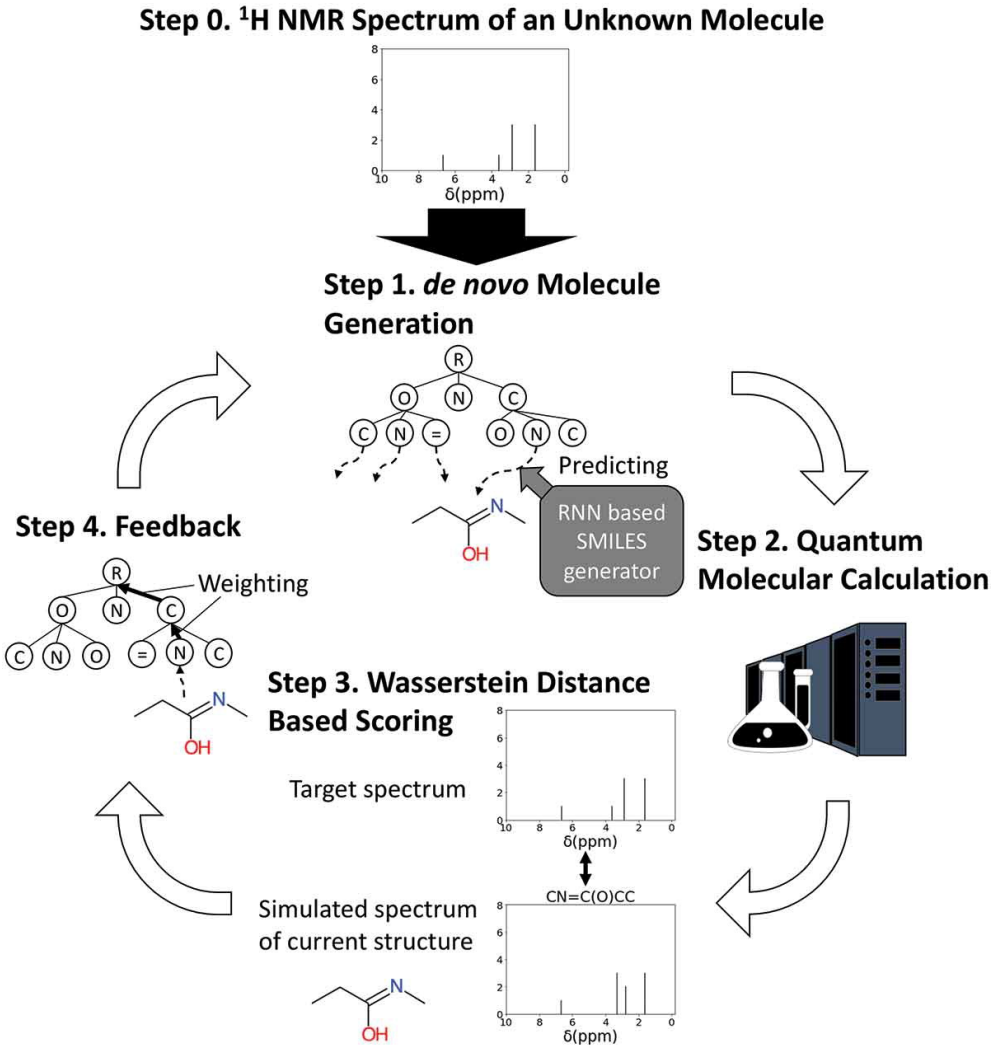

Fig. 1 Concept of NMR-TS

ChemTS

The input for ChemTS is a database of SMILES strings and an evaluation function, which quantifies the goodness of a generated molecule. Starting from a root node that represents the beginning of a SMILES string, the ChemTS algorithm builds a search tree, in which each node corresponds to one SMILES symbol. The ChemTS search process consists of four procedures: selection, expansion, simulation, and backpropagation. In the selection step, the tree is traversed from the root to a leaf by recursively choosing the child node that has the maximum upper confidence bound (UCB) based score at each branch. A path from

the root to the leaf node becomes a SMILES prefix. In the expansion step, several child nodes are added to the leaf node. Upon tree expansion, a selected prefix serves as the input for the RNN pretrained on the database. With the SMILES prefix as an input, the RNN can predict the next symbol after the prefix and elongate the length of the prefix by one. By repeating this elongation step until a terminal symbol appears, a complete SMILES string is generated. The generated molecule is evaluated using the evaluation function and then the tree is updated accordingly during the backpropagation procedure. The input database for pretraining the RNN can be either a general database with no specific molecular characteristics or a specific database containing field-specific SMILES strings.

To perform the massive DFT computations, we parallelized the tree search part of ChemTS using Open MPI based on the virtual loss approach. We used the following scoring in the selection step to avoid concentrating the DFT computations on one node.

$$ucb_i = \frac{S_i}{v_i + w_i} + CP_i \frac{\sqrt{v_p + w_p}}{1 + v_i + w_i}$$

Here, s_i is the total score obtained by node i, v_i is the total visit number of i, w_i is the total virtual visit number of i (virtual loss), v_p is the total visit number of parent node p of i, w_p is the total virtual visit number of p, P_i is the probability of i among the children of p, and C is a constant that controls the exploration–exploitation trade-off.

NMR spectrum prediction

We tested above methods on ^{1}H spectrum of 9 molecules "unknown" by the database. On 6 out of 9 test molecules, our method successfully found the unknown molecular structure based on its spectrum. For the other 3 test molecules, the best candidate found by our methods is at least as good as the best match found in database. Amount all 9 best matches of our methods, none of them has appeared in the training database, which proves the *de novo* molecular generator is not just repeating information learned from training set.

For the computation of ^{1}H-NMR spectra, molecular structures whose atom positions are described in Cartesian coordinates were made by converting a SMILES format to a 3D molecular structure through the function implemented in the RDKit library. We calculated ^{1}H NMR spectra by using density functional theory (DFT) 10 at the B3LYP/3-21G* level on the optimized structure at the universal force field (UFF) level.

Wasserstein Distance and Evaluation function

If we imagine two ^{1}H NMR spectrum as two piles of dirt, the Wasserstein distance is the minimum amount of work needed to reshape one into the other[2]. We use the following Wasserstein distance-based scoring function to evaluate the similarity between two NMR spectrum.

$$Wasserstein\ Score(M_g, M_t) = 1 - tanh(WD(M_g, M_t) + \alpha Penalty(M_g, M_t))$$

$$Penalty(M_g, M_t) = |C(M_g) - C(M_t)| + |H(M_g) - H(M_t)|$$

where $WD(M_g, M_t)$ is the Wasserstein distance between the calculated ^{1}H NMR spectra of generated molecule M_g and target molecule M_t, $C(M)$ and $H(M)$ represent the numbers of carbons and hydrogen in molecule M, and α is a parameter indicating the strength of the penalty. Note that the range of the Wasserstein Score is between 0.0 and 1.0, with 1.0 represents a perfect match between spectra.

Database

We downloaded molecules in form of SMILES with PCCDB-ID from 1 to 138895. We run a selection on these 138895 molecules to eliminate all SMILES contains elements other than C, H, N, O. Also, we have eliminated molecules with chemical charge (such as O+, N+, C-). After selection, 9880 molecules remained

and used as the SMILES database.

Trie enhancement of ChemTS

In the context of ChemTS, the MCTS is essentially executed on a prefix search tree. One advantage of ^{1}H NMR spectrum identification is that an enormous number of molecular spectra have been recorded and stored in databases. An intuitive way of utilizing such information is to preload the MCTS prefix search tree with the SMILES strings of the molecules in the database and update the scoring of each traversed node with the Wasserstein Score between the database spectrum and the target spectrum. We implemented this idea by constructing a trie tree as follows. At every iteration, we inserted one database SMILES string into the trie, on which the nodes were defined in the same way as in ChemTS. After each insertion, the Wasserstein Score of the added SMILES string was used to update the weight of each visited node. The number of preloaded molecules is called the trie size. In our experiments, we tested trie sizes of 0, 1, 100, 1000, and 9800.

Results

We tested NMR-TS on ^{1}H spectrum of 9 molecules "unknown" for the database. On 6 out of 9 test molecules, our method successfully found the unknown molecular structure based on its spectrum. For the other 3 test molecules, the best candidate found by our methods is at least as good as the best match found in database. Amount all 9 best matches of our methods, none of them has appeared in the training database, which proves the *de novo* molecular generator is not just repeating information learned from training set.

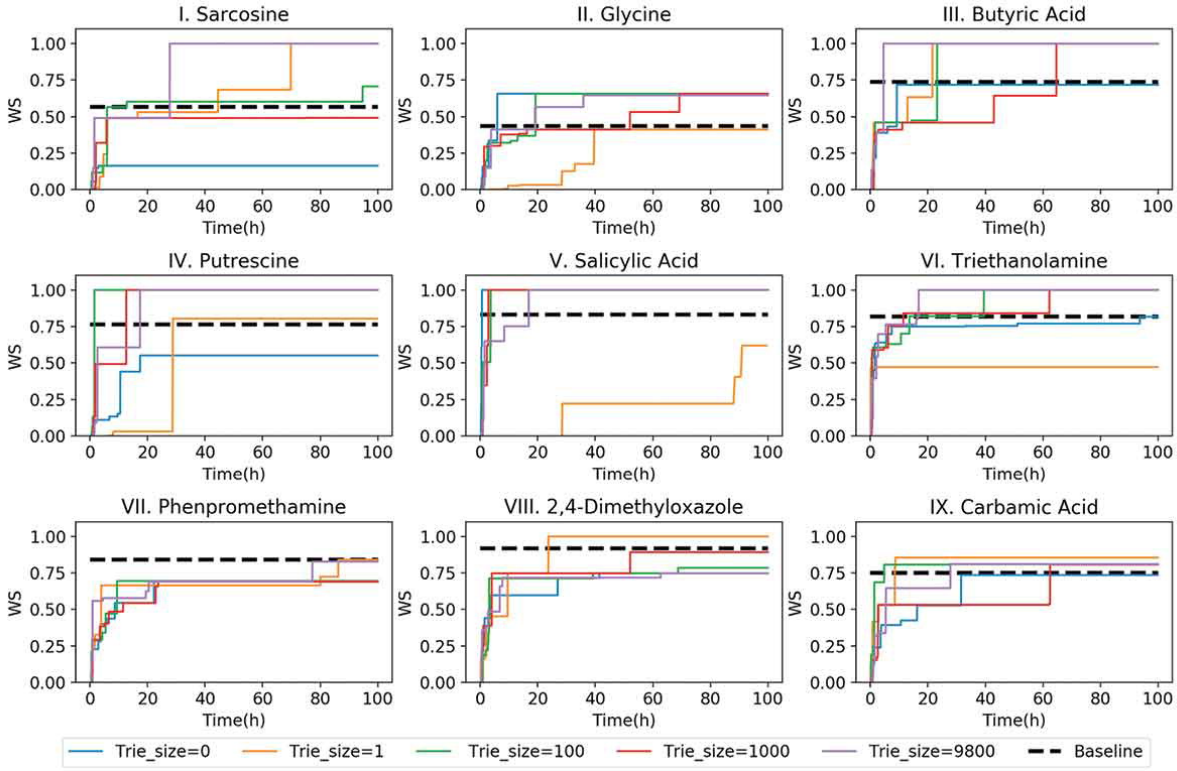

Fig. 2 NMR-TS results under different size of Trie tree for 9 target molecules.

References

1 X. Yang, J. Zhang, K. Yoshizoe, K. Terayama and K. Tsuda, *Sci. Technol. Adv. Mater.*, 2017, **18**, 972–976.

2 A. Ramdas, N. G. Trillos and M. Cuturi, *Entropy*, 2017, **19**, 47.