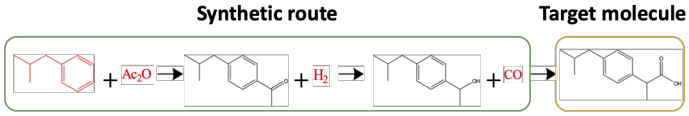
Retrosynthetic analysis using Bayesian inference

Zhongliang Guo
The Graduate University for Advanced Studies, Department of Statistical Science 5-year Doctoral Course – 5rd year

Chemical Synthesis and Retrosynthetic Analysis

Chemical synthesis is the process to obtain a product or multiple product by performing chemical reactions. Given the reactants, hand-coded rules or machine learning model can be used to predict the product(s).



Reaction prediction model using machine learning

	Top 1	Top 3	Top 5
Template-based (Coley et al. 2017)	71.8	86.7	90.8
WLDN (Jin et al. 2017)	79.6	87.7	89.2
Transformer (Schwaller et al. 2018)	90.4	94.6	95.3

Retrosynthetic analysis is the inverse problem of reaction prediction, which is to predict the reactants from the given product. In recent two years, some researches have applied machine learning models to this problem.

Retrosynthesis prediction model using machine learning

	Top1	Top10	Top50
Rule-based (Segler et al. 2017)	31.0	63.3	72.5
Seq2Seq (Liu et al. 2017)	35.4	65.1	69.5
Similarity (Coley et al. 2017)	37.3	74.1	85.3

The inverse problem is harder than the direct problem, especially for retrosynthesis analysis.

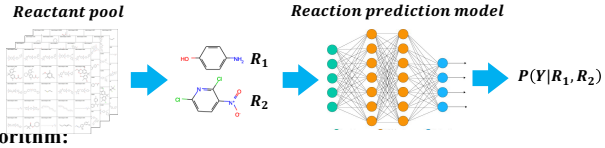
- The solution is not unique. The target product may have several synthesis route.
- The solution space is discrete, we always have to deal with the combinatorial explosion.
- Chemical formula of the product have less information than that of the reactants because we don't know the side product.

Proposed Method for Retrosynthetic Analysis

In this research, we focus on single step retrosynthetic analysis, and propose a Bayesian approach for exploration in solution space.

Outline of this work:

Forward prediction $R_1 + R_2 \rightarrow Y$ (Target product) **Inverse problem** $P(R_1, R_2|Y) \propto P(Y|R_1, R_2) \cdot P(R_1, R_2)$



Algorithm:

Input: K, S, T, M, θ, Y

Output: $\{(r_1, r_2, y)_t^s | \, s = 1, \ldots, S; t = 1, \ldots, T\}$

Set $t = 0$, *reactant pairs* $\{(r_1, r_2)_0^s | \, s = 1, \ldots, S\}$, *evaluate each pair* $(r_1, r_2, y)_0^s = f_Y((r_1, r_2)_0^s)$ *using forward prediction model.*

for t *in* $1, \ldots, T$ *do*

$\{(r_1, r_2)_t^{cand}\} = M\left((r_1, r_2, y, z)_{t-1}^{\{S\}}\right)$ *based on the transition function*

evaluate each pair $(r_1, r_2, z)_t^{cand} = g_\theta^t\left((r_1, r_2)_t^{cand}\right)$ *using simple acquisition function.*

$sort\left(z_t^{\{Cand\}}, order = descend\right)$, *take top* S *pairs*

evaluate each pair $(r_1, r_2, y)_t^s = f_Y((r_1, r_2)_t^s)$

update acquisition function $g_\theta^t \rightarrow g_\theta^{t+1}$ *using* $(r_1, r_2, y)_t^{\{S\}}$

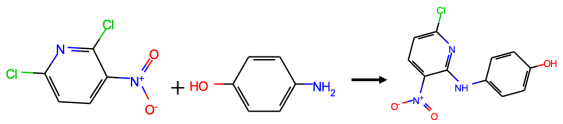
end for

This algorithm can be considered as a lazy Bayesian optimization. Due to the combinatorial explosion, the acquisition function cannot evaluate every point of the solution space. By combining the sequential Monte Carlo (SMC) sampler, we tried to balance exploiting high score reactant pair neighbor and exploring unsearched space.

Experiment and Result

Reaction data

The reaction data used to train the reaction prediction model and comprise the reactant pool is from U.S. patent literature extracted by Lowe[1]. All the reactions are described using SMILES strings.



SMILES Notation: O=[N+]([O-])c1ccc(Cl)nc1Cl.Nc1ccc(O)cc1 (Reactants part)
>>O=[N+]([O-])c1ccc(Cl)nc1Nc1ccc(O)cc1 (product part)

Reaction prediction model

Transformer model[2] gives the state-of-the-art accuracy, so we use this model for reaction prediction. That model was designed for machine translation. Taking reactant SMILES strings as sentence in one language and product SMILES string as sentence in another language, transformer model can be trained by reaction data.

Sequential experiment design approach and stochastic sequential experiment design approach Comparison

Using transformer model as reaction prediction model, it takes 30 sec for 1000 reaction prediction. Since the size of reactant pool is over 600,000, exhausted search for one reactant reaction will take over 5 hours, exhausted search for two reactant reaction will take over 3 *10^6 hours. Even using simple surrogate model such as random forest takes over 3 min for predicting all the 600,000 reactants. So we tried apply a sequential experiment design approach (SEDA) and stochastic sequential experiment design approach (S-SEDA) to reduce the experiment time.

Reaction No.	SEDA first try	SEDA second try	S-SEDA first try	S-SEDA second try
0	5	5	3	10
1	15	6	5	10
2	3	6	2	28
3	10	2	3	10
4	4	6	3	17
5	52	71	15	53
6	10	4	8	2
7	33	24	29	17
8	2	6	31	25
9	5	3	4	4
mean of step No. to reach true reactant	13.9	13.3	10.3	17.6
	SEDA mean elapsed time/step (min)	3.95	S-SEDA mean elapsed time/step	0.295

Table 1: Applied sequential experiment design approach and stochastic sequential experiment design approach to random-selected 10 reactions. Set one reactant as unknow, another one as known. In each step, 100 reactant pairs were evaluated by transformer model. Random forest (RF) was used as acquisition function. In SEDA, RF evaluate all reactants in reactant pool; while in S-SEDA, RF evaluated 1000 reactants proposed by transition function.

Stochastic sequential experiment design approach for reactant pair search

Next, we tried applying stochastic sequential experiment design approach to 2 reactant reactions. The solution space is 3.6 * 10^11. Acquisition function cannot be applied to every pairs.

Reaction No.	First try	Second try	Third try	Mean of step number	Succeed times
0	100	74	65	79.667	2
1	33	13	17	21.000	3
2	45	17	11	24.333	3
3	17	27	16	20.000	3
4	26	11	100	45.667	2
5	38	18	85	47.000	3
6	33	88	47	56.000	3
7	100	88	43	77.000	2
8	100	25	100	75.000	1
9	11	7	10	9.333	3
mean of step No. to reach true reactant	50.3	36.8	49.4	45.5000	25

Table 2: Applied stochastic sequential experiment design approach to random-selected 10 reactions. Reactant pair were unknow. Maximum step size was set to 100. In each step, up to 10000 reactant pairs were proposed by transition function, and evaluated by acquisition function (random forest). Transformer model evaluated 1000 reactant pairs in each step.

Reference

1. Lowe, D. M. *Extraction of Chemical Structures and Reactions from the Literature*. (2012).
2. Schwaller, P. *et al.* Molecular Transformer for Chemical Reaction Prediction and Uncertainty Estimation. (2018). doi:10.26434/chemrxiv.7297379.v1