Vol.2016-BIO-45 No.10
2016/3/19

言語モデルベースの化学構造生成手法の提案と生体活性分子をターゲットにした Inverse-QSAR モデルへの適用

池端久貴[†1]　本郷研太[†2]　磯村哲[†3]　前園涼[†2]　吉田亮[†1†4]

概要：本研究では，統計的言語モデルと逐次モンテカルロ法を組み合わせた分子設計法を提案する．既存化合物から学習された化学構造のパターンを基に単語単位の生成モデルを構成できるため，これまでのフラグメントを用いた手法でカバーしきれなかった多様な化学構造を得ることができる．本手法を用いて，高い生体活性を示す医薬品候補の生成を試みる．

キーワード：Inverse-QSAR，生理活性分子探索，統計的言語モデル，逐次モンテカルロ法

1. はじめに

医薬品を含む多くの化学製品の開発現場において，計算機を利用した分子設計は非常に有用な手法である．特性から化学構造の予測モデルである Inverse-QSAR は，QSAR に比べると研究成果の報告件数はかなり少ない．その理由として，広大な化学構造空間の取り扱いが難しいことが挙げられる．例えば，元素を C、N、O、S に限定し，非水素原子の総数を 30 までとした場合でも，可能な有機分子の組み合わせ総数は 10^{60} 個に膨れ上がる．

先行研究として，最も良く使われている手法は，遺伝的アルゴリズムに基づき[1]，化学構造を断片化して得られたフラグメントを構成単位として用いる．より多くの化学構造を考慮する場合、必然的に扱うフラグメント種も多くなり，計算時間に影響を与える．

本研究では，統計的言語モデルに基づく化学構造モデルを導入した Inverse-QSAR モデルと逐次モンテカルロ法を用いた分子設計法を提案する．本手法では化学構造の表現にフラグメントを使わずに，言語モデルを用い，出現パターンを既存化合物から事前学習することで、より多様な化合物らしい構造を短時間で生成することを狙う．また、逐次モンテカルロ法と組み合わせることで，Inverse-QSAR モデルにおける条件付分布からの化学構造生成が可能となる．

本発表ではデモとして，生体活性データに関する QSAR モデルを PubChem に登録されている生体活性データから学習，化学構造モデルを PubChem の化学構造データから学習し，高い生体活性を示す医薬品候補の生成を試みる．

2. 手法

2.1 n-gram モデルによる化学構造のモデリング

SMILES 形式は文字列で化学構造を表現するために広く用いられている形式の一つである．各元素は標準的な略記形が用いられ，有機分子骨格を形成する B, C, N, O, P, S, F, Cl, Br，I 以外は[Fe]のように角括弧を用いる.環構造は始点と終点を数字で，側鎖部分は丸括弧で囲むことで表現される．この SMILES 形式で表現された化学構造を n-gram モデルを仮定し学習することで，化学構造の特徴的なパターンを抽出し，化学構造のランダムサンプリングが可能となる．ただし，化合物の環構造，側鎖などの分岐構造を文字列で表現する際には，離れた文字間の依存関係が現れるため，以下のような工夫を施す．

- 開始している環（ペアが現れていない数字）の個数，開始している側鎖の有無に応じて，部分文字列を異なるクラスに分類し，それぞれのクラス毎に学習を行う
- 閉じた側鎖の情報を削減するオペレータ Ψ の導入
  (例) Ψ(CCC(CCCC)O)=CCC(C)O

以上を考慮した上で，SMILES 文字列を$\boldsymbol{s} = (s_1, \cdots, s_{|\boldsymbol{s}|})$，a 文字目から b 文字目までの部分文字列を$\boldsymbol{s}_{[a:b]}$とするとき，改良 n-gram モデルは以下のように表すことができる．

$$p_n(\boldsymbol{s}) = \prod_{i=1}^{|\boldsymbol{s}|} \sum_{k=1}^{|A|} {p_n}^{(k)} \left(s_i \mid \Psi(\boldsymbol{s}_{[i-n+1:i-1]})\right) I(\boldsymbol{s}_{[1:i-1]} \in A_k),$$

ここで${p_n}^{(k)}$はクラスA_kにおける文字生成確率を表し，$|A|$個あるグループへの割り当ては，生成文字以前の部分文字列の非ペアの数字の個数，閉じていない側鎖の有無から非確率的に決定される．パラメータ推定は既存化学構造から Kneser-Ney smoothing などの推定方法を用いる．

2.2 Inverse-QSAR モデリング

目標の特性y^*が与えられたときに，その構造$\boldsymbol{s}$についての条件付き分布$p(\boldsymbol{s} \mid y^*)$を得たい．ベイズの定理を用いると$p(\boldsymbol{s} \mid y^*) \propto p(y^* \mid \boldsymbol{s}) p(\boldsymbol{s})$となり，右辺第一項は QSAR モデルであり，第二項は構造に関する事前知識となり，2.1 節で提案されたモデルを用いる．QSAR モデルを構築する際には，化学構造をフィンガープリントと呼ばれるベクトル表現されたものを入力とすることが多い．本稿では適当な記述子$\phi: \mathbf{s} \to \mathbf{x}$を用いて p 次元のベクトル$\mathbf{x}$が得られたとする．目的の特性値 Y は二値変数とし，未知の係数ベクトル$\mathbf{w}$を用いて，$P(Y = y \mid \mathbf{x}, \mathbf{w}) = \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x})^y \left(1 - \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x})\right)^{1-y}$で表されるロジスティックモデルを仮定する．ここで $\sigma(x) =$

†1 総合研究大学院大学統計科学専攻
†2 北陸先端科学技術大学院大学
†3(株)地球快適化インスティテュート
†4 統計数理研究所

$\{1+\exp(-x)\}^{-1}$ である．データ$\{\mathbf{x}_i, y_i\}_{i=1,\cdots,N}$が観測された際の負の対数尤度は，

$$\mathrm{NLL}(\mathbf{w}) = -\sum_{i=1}^{N} y_i \log\sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i) + (1-y_i)\log\sigma(1-\mathbf{w}^{\mathrm{T}}\mathbf{x}_i)$$

となる． $\mathbf{w}$ の推定は反復再重み付け最小二乗法などを用いて NLL を最大にする $\hat{\mathbf{w}}$ を求める．

2.3 逐次モンテカルロ法による化学構造生成器

得られた $\boldsymbol{s}$ の事後分布を目標分布とし，逐次モンテカルロ法[2]を適用する．ステップ t における目標分布γ_tを

$$\gamma_t(\boldsymbol{s}) = \{P(Y=1 \mid \boldsymbol{\phi}(\boldsymbol{s}), \hat{\mathbf{w}})\}^{\kappa(t)}\hat{p}_n(\boldsymbol{s})$$

とする．ここで，$\kappa(t)$は正の単調関数で少なくとも最終ステップにおいては$\kappa(t)=1$ となる関数である．逐次モンテカルロ法の手続きとしては以下のようになる．

1. 初期化：初期値 ${\boldsymbol{s}_1}^{(i)}, {W_1}^{(i)}$ を適当に与える．$(i = 1,\cdots N)$
2. 時刻 t において： ${\boldsymbol{s}_{t-1}}^{(i)}$から${\boldsymbol{s}_t}^{(i)}$への局所遷移を以下の手続きで行う $(i = 1,\cdots N)$.
   - u~Binom(L,0.5)，ここで L は事前に決定された定数．
   - 右端から u 文字縮め，右端から L-u 文字生成する．このときの生成は 2.1 で提案した生成モデル $\hat{p}_n(\boldsymbol{s})$ を用いる．
3. 重みを以下のように更新する．
$${w_t}^{(i)} := {w_{t-1}}^{(i)} \frac{\{P(Y=1 \mid \boldsymbol{\phi}({\boldsymbol{s}_t}^{(i)}), \hat{\mathbf{w}})\}^{\kappa(t)}}{\{P(Y=1 \mid \boldsymbol{\phi}({\boldsymbol{s}_{t-1}}^{(i)}), \hat{\mathbf{w}})\}^{\kappa(t-1)}}, \text{for } i = 1,\cdots N$$
4. 重みを正規化し（総和で割る）， ${W_t}^{(i)}$を得る．
5. $\left(\frac{1}{N}\sum_{i=1}^{N} {W_t}^{(i)2}\right)^{-1} \leq {}^{N}/_{2}$ ならば，正規化された重みをパラメータとする多項分布を用いてリサンプルを行う．重みを${W_t}^{(i)} = \frac{1}{N}$ とする．
6. OpenBabel を用いて${\boldsymbol{s}_t}^{(i)}$の並びを変える．その際，先頭原子はランダムに指定する．
7. t:=t+1 とし，2 に戻る．

3. 結果と考察

QSAR 関数の学習は，PubChem の bioassay データを用いて行った[3]．NCI human tumor cell line growth inhibition assay データには活性分子が 2104，不活性分子が 38796 あり，不活性分子数が活性分子数を同数になるようにランダムにサンプルを選んだ．その上で，訓練データとテストデータを 1:1 に分け，R の glmnet パッケージを用いてパラメータ推定を行った．フィンガープリントは OpenBabel の FP2 を用いた。結果，テストデータセットにおける判別精度は 0.824 となった（図１）．

化学構造の学習は PubChem 登録されている化学構造をランダムに 50000 個選び，10-gram モデルをベースにした提案モデル(2.1 参照)を用いて行った．

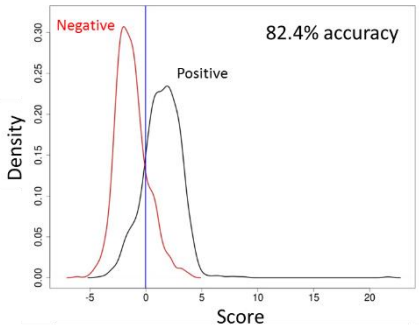

図１ テストデータの活性とスコアの同時分布

次に，500 個の粒子を用いて逐次モンテカルロ法を 25step 行い，活性を示す化学構造の生成を行った．図 2 では，逐次モンテカルロの各ステップにおけるスコア$\hat{\mathbf{w}}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{s})$の分布をボックスプロットで示している．PubChem からランダムに選ばれた初期構造においては，スコアの小さな構造が大半を占めていたが，ステップが進むにつれ，スコアの大きな化学構造が出現しており，スコアの値から，これらは有望な活性分子の候補と言える．

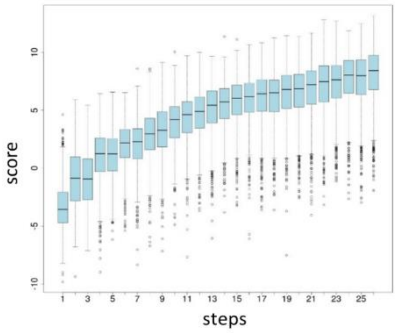

図２ 逐次モンテカルロにおけるスコアの推移

図３に生成器で得られた高いスコアを持つ化学構造の一例を示す．

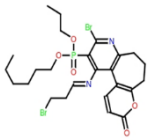

図３ 生成された化学構造の一例

参考文献

[1]E. Lameijer *et.al*, The Molecule Evoluator. An Interactive Evolutionary Algorithm for the Design of Drug-Like Molecules, *J. Chem. Inf. Model.*, 46, 545-552, 2006.

[2]P. Del Moral *et.al*, Sequential Monte Carlo samplers, *J. R. Statist. Soc. B*, 68, 411-436, 2006.

[3]Kim S *et.al*, PubChem Substance and Compound databases. *Nucleic Acids Res*. 2015 Sep 22. pii: gkv951.