Vol.2017-MPS-113 No.32
Vol.2017-BIO-50 No.32
2017/6/24

機械学習を用いた 創薬におけるリード化合物最適化経路の予測

安尾 信明[1,a)]　渡辺 敬介[2]　新井 直樹[1]　関嶋 政和[1,3]

概要：新薬開発における段階の一つであるリード最適化では，化合物は標的蛋白質への活性だけでなく脂溶性や毒性など様々な性質を考慮した合成展開がなされ，その薬としての性質を向上させていく．本研究の目的は，製薬会社内における過去の化合物最適化経路のデータを利用して，新規の標的に対するリード最適化を自動的に行うシステムを作成することである．本発表では，このシステムにおける化合物の評価関数に相当する化合物の薬らしさの予測を行うため，機械学習を用いて最適化過程で化合物が合成された順序の予測を試みた．

キーワード：リード最適化，機械学習，創薬 ，ランク学習

Yasuo Nobuaki[1,a)]　Watanabe Keisuke[2]　Arai Naoki[1]　Sekijima Masakazu[1,3]

1. 序論

我々の生活において薬剤はなくてはならないものであり，新規の薬剤を創出する創薬もまた重要な産業である．しかし，新規な薬剤を開発するためのコストは年々増加しており [1]，情報技術を用いたコストダウン手法の需要が高まっている．創薬プロセスの一部であるリード最適化は，特に低分子の新規薬候補化合物を創出する際に重要なプロセスである．リード最適化では，既に薬剤標的に対してある程度の活性をもつ化合物を，より薬として好ましい性質を持つ化合物に変化させる [2]．具体的には，化合物の構造を変化させることで，標的に対する活性や，ADMET(absortion, distribution, metabolism, excretion, toxicity) を始めとする化合物の物理化学的特性といった様々な性質について，それらを総合的に最適化していく過程である（図 1）．

[1] によれば，リード最適化には創薬全体の約 17% の費用が掛かっている．リード最適化のコストが高い原因は，

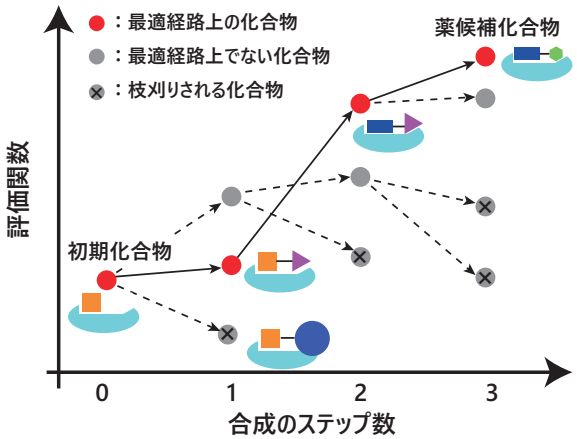

図 1　リード最適化の模式図

薬らしい化合物を発見するために多くの化合物が合成され，そのうちのわずかしか薬らしい化合物が得られないためである．したがって，化合物最適化の経路を過去のデータを用いて予測し，実際に合成する化合物数を削減することが可能となれば，創薬のコストを大きく削減できる可能性がある．

本研究の最終的な目標は，化合物を入力とし，最適化後の化合物を出力とするような自動化合物最適化システム（図 2）を作成することにある．このシステムは大きく分けて二つの要素からなる．ひとつは最適化中の候補化合物の探索であり，もう一つは探索された候補化合物の評価で

[1] 東京工業大学 情報理工学院 情報工学系
Department of Computer Science, Tokyo Institute of Technology
[2] 東京工業大学 工学部 情報工学科
Department of Computer Science, Tokyo Institute of Technology
[3] 東京工業大学 科学技術創成研究院 スマート創薬研究ユニット
Advanced Computational Drug Discovery Unit, Tokyo Institute of Technology
[a)] yasuo.n.aa@m.titech.ac.jp

Vol.2017-MPS-113 No.32
Vol.2017-BIO-50 No.32
2017/6/24

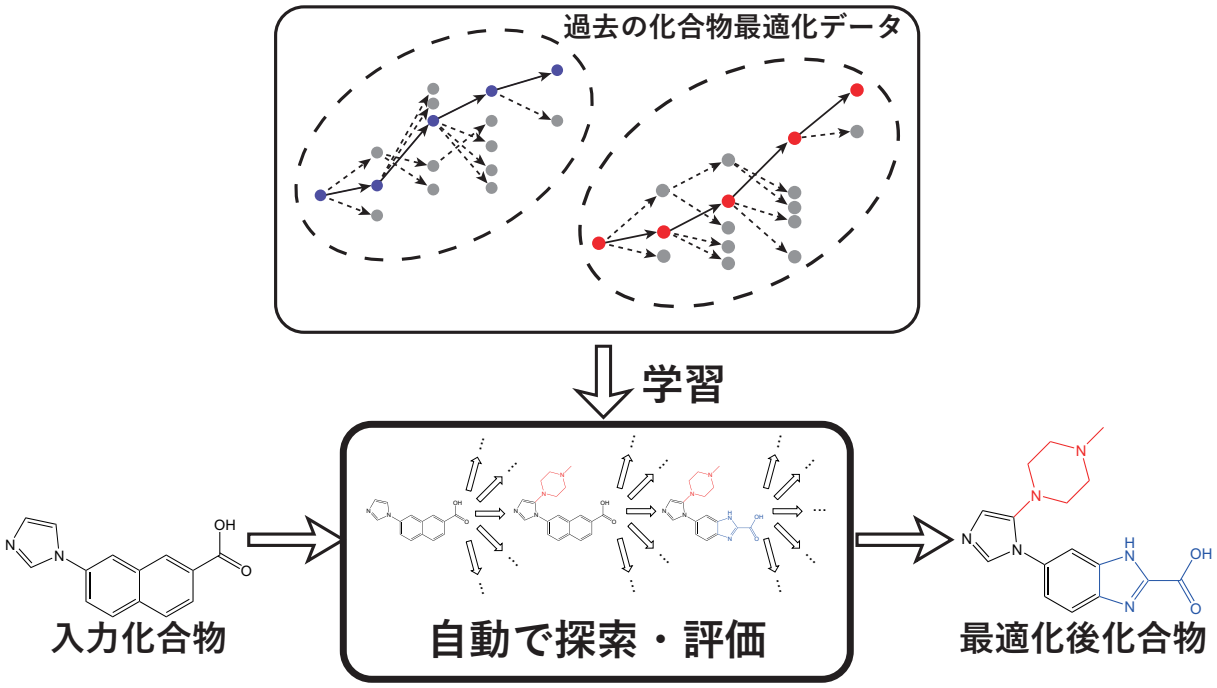

図 2 自動化合物最適化システムの概念図

ある．前者は matched molecular pair (MMP) を用いた化合物の置換を用いて達成可能である．MMP はある一部の構造や官能基だけが異なる二つの化合物の組のことを表し [3]，MMP を用いた化合物最適化に関する研究は既に行われている [4], [5].

後者については，定量的構造活性相関 (quantitative structure-activity relationship: QSAR) や定量的構造物性相関 (quantitative structure-property relationship: QSPR) といった化合物の構造から標的への活性や特定の物性を予測する試みが広く行われている [6], [7] が，リード最適化においては複数の活性や物性を同時に最適化しなければならず，統一された指標があることが望ましい．また，化合物の「薬らしさ」を表す値としてリピンスキーの rule of 5[8] や QED[9] が存在するが，これらは化合物の官能基の情報や物理化学的な特徴を用いて設計された量であり，創薬化学者の間で広く用いられているが，リード最適化にはこの他様々な考慮すべき特徴が存在するため，これだけでは十分に化合物最適化の予測が可能であるとはいえない．

本研究では，製薬会社内に存在する過去のリード最適化研究で合成された化合物を特徴とし，学習することにより化合物の薬らしさを予測する．リード最適化ではより後に合成された化合物の方がより薬らしい化合物であると考えられるため，化合物の合成順序を予測することとし，複数のランク学習手法を適用してその精度を比較・検討した．ランク学習は情報検索などの分野で多く用いられる手法で，データの順序を予測する手法である [10]．ランク学習には pointwise, pairwise, listwise の 3 手法があるが，本研究で用いているのは pointwise 手法と pairwise 手法である．Pointwise 手法は単純な回帰問題と同様であり，予測対象一つを一つの値と対応付け，その値の比較によって順序の予測を行う．Pairwise 手法は予測対象二つの組について，いずれの順序が先かを予測する分類問題である．Listwise 手法は予測対称のリストについて，直接全ての順序を予測する手法である．本研究では，Pointwise 手法と Pairwise 手法について比較検討を行った．

2. 手法

2.1 データセット

本研究におけるデータセットは，武田薬品工業株式会社より提供された 17 種の標的タンパク質に関するデータセットである．ある薬候補化合物の最適化過程において合成された化合物が，合成された時系列によって並んだデータをプロジェクトと呼ぶ．それぞれの標的タンパク質はすべて異なり，タンパク質の分類上二つの大きなグループ A, B に分けられる．プロジェクト 1–6 はグループ A, 7–17 はグループ B に属する．各プロジェクトの化合物数は，最大 570, 最小 291, 平均 484 である．

各化合物のラベルは合成順序とした．化合物は時系列順に番号が振られており，回帰モデルについてはこの番号を $[0, 1]$ に正規化したもの，ランク学習モデルについてはこの順序をそのままラベルとしており，ランク学習モデルについては同じタンパク質内の組のみを比較する．

2.2 特徴量

プロジェクトに含まれる化合物は ECFP6[11] によって特

表 1 使用手法

手法名	種別	詳細	パラメータ
SVM (linear)	pointwise	SVM 線形カーネル (回帰モデル)	$C : 1.0 \times 10^{-5}$
SVM (rbf)	pointwise	SVM RBF カーネル (回帰モデル)	$C : 1.0, \gamma : 0.1$
Random forest	pointwise	ランダムフォレスト	特徴量数 : 25
rankSVM	pairwise	SVM 線形カーネル (判別モデル)	$C : 1.0$
logistic	pairwise	ロジスティック判別モデル	$\alpha : 1.0 \times 10^{2}$
lasso	pairwise	L1 正則化付き最小二乗法	$\alpha : 1.0 \times 10^{-3}$

徴ベクトル化されている．ECFP (Extended-connectivity fingerprint) は部分構造ベースの fingerprint であり，以下のアルゴリズムにより計算される．まず，化合物中の各原子に ID を割り当てる．次に，各原子について，対象の原子からの距離 0, 1, 2, ... 以内にある原子からなる部分構造を，原子の ID とともに Morgan 法によりハッシュ化し，そのハッシュ値を規定のビット数で割った剰余の位置のビットを立てる．ECFP6 では，対象の原子からの距離が 6 以下の部分構造を扱う．なお，本研究で提供されているデータセットは 512bit であるが，化合物の化学構造が提供されておらず，ECFP の特徴量の順序が入れ替えられているため，特徴量を利用した学習は可能であるが，元の化合物が特定できないようになっている．

各特徴について，すべて 0 もしくはすべて 1 であるような特徴は存在しなかった．また，相関係数が 0.95 を超えるような特徴量のペアは存在しなかった．

2.3 学習手法

本研究ではランク学習の pointwise および pairwise モデルを検討し，その性能を比較した．使用した手法，pointwise/pairwise の種別, 詳細, 最適化後のパラメータを表 1 に示す．実装には python3.5.2, scikit-learn version 0.18.1[12] を使用した．

Pairwise 手法については，データを i, j, 特徴量を x, ラベルを y としたとき，(x_i, x_j, y_i, y_j) のランク予測問題は，線形モデルの場合に限り $(x', y') = (x_i - x_j, sign(y_i - y_j)$ の判別問題として解くことができることを利用し，線形の判別機を用いている．学習器のハイパーパラメータは，同時に提供された別のタンパク質と化合物に関するデータを用いて最適化した．pairwise 手法については，データ数が増加するため確率的勾配降下法を用いたミニバッチ学習 [13] を行った．ミニバッチのデータ数は 500, epoch 数は 170 である．

学習結果の評価は Spearman の順位相関係数 ρ を用いた．順位相関係数は $-1 \leq \rho \leq 1$ で，一様にランダムな予測における ρ の平均値は 0 となる．また，それぞれについて，相関が存在するかについて有意水準 0.05 で検定を行った．

表 2 予測精度

手法名	種別	平均の順位相関係数 ρ
SVM (linear)	pointwise	0.321
SVM (rbf)	pointwise	0.344
Random forest	pointwise	0.301
rankSVM	pairwise	0.238
logistic	pairwise	0.262
lasso	pairwise	-0.041

3. 結果・考察

17 プロジェクトの結果の平均を表 2 に示す．6 手法のうち，lasso のみうまく学習できていない結果となっているが，その他はすべて相関係数が 0.2 以上となり，いずれも p¡0.05 で有意に正の相関があり，ランダムな予測に比べて精度よく予測できていた．lasso がうまく学習できていない理由については，今回の学習における特徴ベクトルがスパースでないため，特徴選択が悪影響を及ぼしてしまった可能性が考えられる．また，全体として pointwise では pairwise に比べて精度よく予測できていたことについて，後述するプロジェクト 1 と 6 など，異なるプロジェクト間の化合物の比較も予測に寄与している可能性が考えられる．

次に，pointwise, pairwise でそれぞれ最も精度のよかった SVM (rbf), logistic に加え学習がうまく行かなかったと思われる lasso について，各プロジェクトの順位相関係数を図 3 に示す．学習がうまく行った SVM (rbf), logistic については，プロジェクト 1, 6, 9, 12, 17 などで高い精度を得られている．これらのプロジェクトでは，過去のデータから十分に情報を得られていると考えられる．一方，プロジェクト 4, 11, 13 などではどの予測手法も十分な情報が得られていない．これは，他のプロジェクトと化合物の最適化戦略が異なることを示していると考えられる．

また，全てのプロジェクトの化合物について，化合物間の類似度を表す Tanimoto 係数を計算したヒートマップを図 4 に示す．黒は類似性が低く，赤，白となるにつれて類似性が高いことを表す．この図より，同一プロジェクト内の化合物間の類似性は比較的高いが，プロジェクト間にまたがって類似する化合物はプロジェクト 1 と 6 といった例外を覗いてほぼ存在しないことが分かる．また，例外的に化合物が類似しているプロジェクト 1 と 6 はそれぞれ予測精度が比較的高いプロジェクトであることもわかる．この

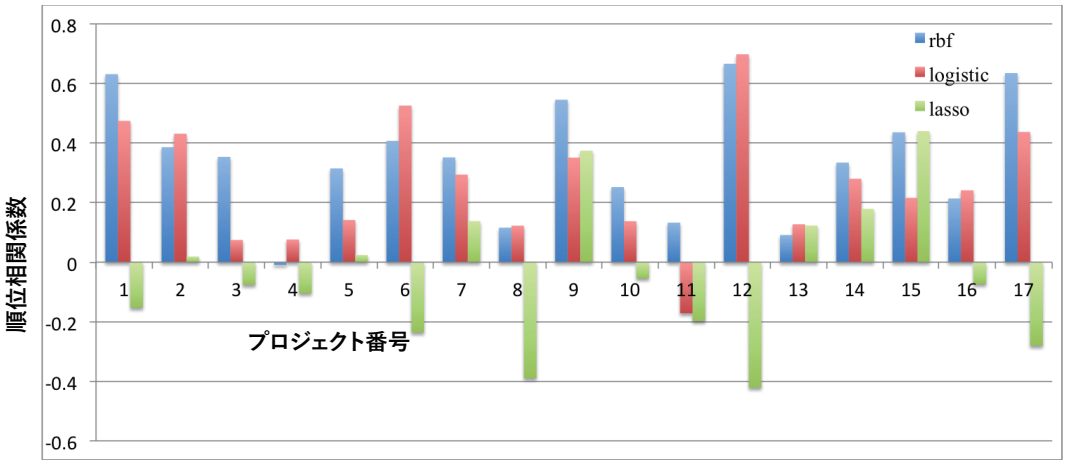

図 3 学習結果．縦軸：順位相関係数，横軸：プロジェクト番号

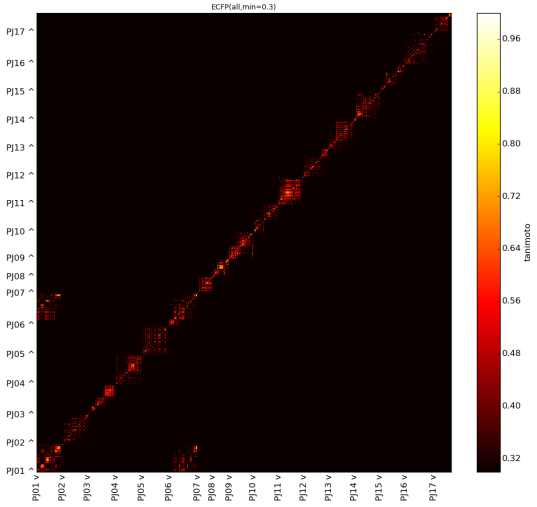

図 4 化合物間類似度のヒートマップ

ことから，類似の化合物を用いているプロジェクトは比較的容易に合成順序を予測できること，そして単純に特徴量が類似しているわけではない化合物の情報だけでも，化合物の合成順序がある程度予測できることがわかる．

さらに，線形モデルにおける各特徴量の重みについて，半数以上の特徴量で重みの値が正になっていることを確認した．これは，全体として特徴ベクトル中で 1 となっているビットの数が多い化合物の方がより後に合成されたと予測していることを示している．ECFP6 における 1 となっているビットの数は化合物の構造の複雑さに対応しており，合成展開が進むに従って化合物がより複雑になっていくことと対応していることが明らかとなった．

4. 結論

本研究では，創薬における化合物最適化を自動的に行うシステムを作成することを目指し，化合物の薬らしさを他の創薬研究における最適化事例から予測する手法について提案・検討した．本研究により，16 プロジェクトの結果を用いることによって，新規のプロジェクトの化合物合成順序をある程度予測できることを示した．また，類似している化合物を含むプロジェクトの合成順序は比較的高精度に予測できること，そして単純に化合物が類似しているわけではないプロジェクトの情報から，化合物の合成順序がある程度予測できることを示している．今後の展望としては，さらに精度を高めるためにどのような情報を付与すればよいか，また，現在の予測手法に化合物の探索手法を組み合わせ，化合物自動最適化システムを構成した実験を行うことなどが挙げられる．

5. 謝辞

本研究の一部は，JSPS 科研費 基盤 (B) 15H02776, 特別研究員奨励費 16J09021 及び JST リサーチコンプレックス推進プログラム「世界に誇る社会システムと技術の革新で新産業を創る Wellbeing Research Campus “ Tonomachi”」, AMED 創薬等ライフサイエンス研究支援基盤事業の支援を受けて行われた．また，本研究で用いたリード最適化に関するデータは，武田薬品工業株式会社から提供された．

参考文献

[1] Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R. and Schacht, A. L.: How to improve R&D productivity: the pharmaceutical industry’s grand challenge, *Nature reviews Drug discovery*, Vol. 9, No. 3, pp. 203–214 (2010).

[2] Keserü, G. M. and Makara, G. M.: The influence of lead discovery strategies on the properties of drug candidates, *nature reviews Drug Discovery*, Vol. 8, No. 3, pp. 203–212 (2009).

[3] Tyrchan, C. and Evertsson, E.: Matched Molecular Pair Analysis in Short: Algorithms, Applications and Limitations, *Computational and Structural Biotechnology Journal*, Vol. 15, pp. 86–90 (2016).

[4] Ritchie, T. J. and Macdonald, S. J.: Heterocyclic replacements for benzene: maximising ADME benefits by considering individual ring isomers, *European Journal of Medicinal Chemistry*, Vol. 124, pp. 1057–1068 (2016).

[5] Weber, J., Achenbach, J., Moser, D. and Proschak, E.: VAMMPIRE: a matched molecular pairs database for structure-based drug design and optimization, *Journal of medicinal chemistry*, Vol. 56, No. 12, pp. 5203–5207

(2013).
[6] Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., Dearden, J., Gramatica, P., Martin, Y. C., Todeschini, R. et al.: QSAR modeling: where have you been? Where are you going to?, *Journal of medicinal chemistry*, Vol. 57, No. 12, pp. 4977–5010 (2014).
[7] Zheng, W. and Tropsha, A.: Novel variable selection quantitative structure- property relationship approach based on the k-nearest-neighbor principle, *Journal of chemical information and computer sciences*, Vol. 40, No. 1, pp. 185–194 (2000).
[8] Lipinski, C. A., Lombardo, F., Dominy, B. W. and Feeney, P. J.: Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, *Advanced drug delivery reviews*, Vol. 23, No. 1-3, pp. 3–25 (1997).
[9] Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S. and Hopkins, A. L.: Quantifying the chemical beauty of drugs, *Nature chemistry*, Vol. 4, No. 2, pp. 90–98 (2012).
[10] Liu, T.-Y.: Learning to Rank for Information Retrieval, Foundations and Trends in Information Retrieval, Vol. 3, No. 3, pp. 225–331 (online), DOI: 10.1561/1500000016 (2009).
[11] Rogers, D. and Hahn, M.: Extended-connectivity fingerprints, *Journal of chemical information and modeling*, Vol. 50, No. 5, pp. 742–754 (2010).
[12] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E.: Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, Vol. 12, pp. 2825–2830 (2011).
[13] Bottou, L.: Large-scale machine learning with stochastic gradient descent, *Proceedings of COMPSTAT'2010*, Springer, pp. 177–186 (2010).