Vol.2022-BIO-69 No.12
2022/3/10

エンハンサー・プロモーター間相互作用予測問題に対する負例生成手法の提案

古賀 吏[1,a)]　丸山 修[2,b)]

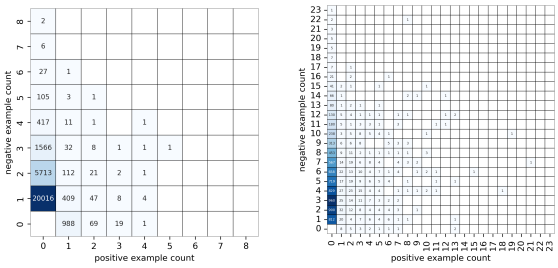
概要：エンハンサー・プロモーター間相互作用は，遺伝子制御において重要なメカニズムである．先行研究の TargetFinder や EP2vec の学習データを染色体単位で分割した交差検証では学習が全くうまく行かないことが報告されており，その原因は TargetFinder の負例の生成方法にあることが指摘されている．本研究では，エンハンサー・プロモーター間相互作用予測問題に対する新しい負例生成手法を提案する．具体的には，与えられた正例の相互作用集合から構成される有向グラフに対する最大フロー問題に還元することにより適切な負例を生成する．生成された負例のほとんどは，本研究で課した制約（正例における各エンハンサーと各プロモーターの出現回数と同程度になること）を満たしていることが分かった．さらに生成された負例を用いた学習では，明らかに予測精度が改善されていることが判明した．

1. はじめに

細胞がその個体に必要なタンパク質を合成するとき，まず DNA 上の遺伝子領域が mRNA に転写され，さらに mRNA がタンパク質に翻訳される．遺伝子領域の転写は，エンハンサーやプロモーターと呼ばれる DNA 領域が互いに相互作用することで開始される．

近年の研究により，幾多の基礎疾患に関連する一塩基多型（SNPs）が，細胞株において特異的なエンハンサー内に存在していることがゲノムワイド関連解析（GWAS）によって確認されている [1], [2]．エンハンサーの SNP による遺伝子発現や基礎疾患の影響の認識は困難であるが [3]，エンハンサーとそれが相互作用するプロモーターが既知であれば，遺伝子発現や疾患への影響の認識が可能となる．すなわち，エンハンサーとプロモーターの相互作用を正確に同定することは，遺伝子発現，細胞分化，及び疾患メカニズムを明らかにするために重要である．

エンハンサー・プロモーター間相互作用予測問題に対する先行研究モデルである TargetFinder[4] は，エンハンサーとプロモーターのそれぞれのヒストン修飾などの状態から得られる特徴量を用いた相互作用予測モデルを提案している．これの学習のために，Hi-C[5] によって得られるエンハンサー・プロモーター間相互作用データから正例相互作用データと，自身で定義した負例相互作用データを生成している．そしてその交差検証の結果，高いパフォーマンスが得られたことを報告している．



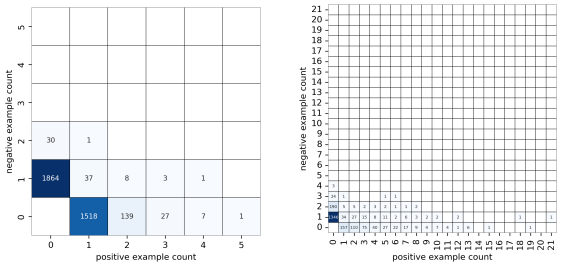
(a)　(b)

図 1: TargetFinder が使用した正負例データセットでの (a) エンハンサーと (b) プロモーターの正例での出現回数と負例での出現回数の組の頻度情報．

さらに，同じ正負例データを用いた手法 EP2vec[6], SPEID[7]，EPIVAN[8]，EPIHC[9] はさらに高い予測精度の結果を報告している．

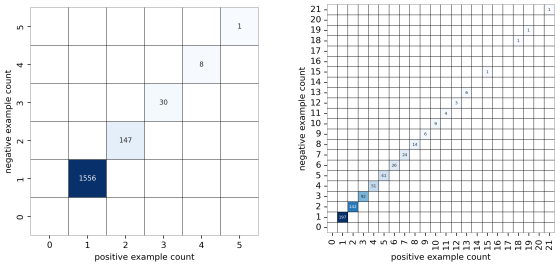
しかしながら，TargetFinder の正負例データセットでは，非相互作用（負例）データの生成方手法に問題があることが指摘されている [10], [11]．具体的には，各エンハンサーと各プロモーターの正例と負例での出現回数に大きな偏りがある（図 1）ことで，交差検証による本質的な学習ができない点である．いまのところ我々が知る限りでは，TargetFinder の負例データを改善する研究報告は成されてない．

[1] 九州大学芸術工学部
School of Design, Kyushu Uniersity
[2] 九州大学大学院芸術工学研究院
Faculty of Design, Kyushu Uniersity
[a)] ylwrvr.t.koga@gmail.com
[b)] maruyama@design.kyushu-u.ac.jp



(a) (b)

図 2: EP2vec が使用した正負例データセットでの (a) エンハンサーと (b) プロモーターの正例での出現回数と負例での出現回数の組の頻度情報.



(a) (b)

図 3: 細胞株を K562 とした時の我々が構成した正負例データセットでの (a) エンハンサーと (b) プロモーターの正例での出現回数と負例での出現回数の組の頻度情報.

エンハンサー・プロモーター間相互作用問題に限らず，一般的には，2 つのオブジェクト間の相互作用問題用の負例データの生成法は自明ではない．しかし，少なくとも，エンハンサー・プロモーター相互作用問題に対する負例データが満たすべき条件の一つは，個々のエンハンサーとプロモーターの正例相互作用と負例相互作用での出現回数が同程度になることだと考えられる．そこで本研究では各エンハンサーと各プロモーターの正例と負例での出現回数が同程度になる制約を課した負例生成手法を提案した．

2. 手法

上記の制約を課した負例生成問題を最大フロー問題 [12] へ還元することで負例の相互作用データを生成する．最大フロー問題は線型計画問題の一つであり，各有向辺に容量が定められている有向グラフ上で始点となる頂点から終点となる頂点まで流すことができる量の最大値を求める問題である．与えられた正例のエンハンサー・プロモーター間相互作用集合から構成される有向グラフに対する最大フロー問題を解くことで，各エンハンサーと各プロモーターの正例での出現回数を超えない最大数の負例を生成することができる．

3. 評価

提案手法で生成した負例の有効性を TargetFinder との比較で解析する．具体的には，本研究で構成した正負例データセットと TargetFinder の正負例データセットと EP2vec の正負例データセットのそれぞれを用いて EP2vec 予測モデルを交差検証で学習し，予測精度を比較する．EP2vec の正負例データセットは，TargetFinder 正負例データセットの負例数を正例数と同数にしたものであり，TargetFinder の負例の問題点である「各エンハンサーと各プロモーターの正例と負例での出現回数の偏り」は残ったままである (図 2 を参照)．

EP2vec はエンハンサーとプロモーターの塩基配列のみを入力として各塩基配列を自然言語処理モデル (doc2vec[13]) を用いて埋め込みベクトルに変換し，その埋め込みベクトルで分類器学習を行う予測モデルである．

分類器には，GBRT (Gradient Boosting Decision Tree) と k-NN (k-Nearest Neighbor) を採用した．GBRT は複数の弱学習器である決定木によって構成される分類器であり，木の深さ $D = 25$ を固定して木の数 $N = 100, 1000, 4000$ のそれぞれで予測を行う．k-NN は近傍数を k とした時，与えられた埋め込みベクトルのベクトル空間中の位置から距離が最も近い k 個の訓練例のクラス・ラベルの多数決で分類を行う手法であり，$k = 5, 10, 15$ のそれぞれで予測を行う．

交差検証では，データセットをランダムに訓練用とテスト用に分割する一般的な交差検証と，染色体単位で訓練用とテスト用に分割する交差検証の 2 通りを行う．

4. 結果

図 3 は，細胞株 K562 に対する，各エンハンサーとプロモーターのそれぞれの正例での出現回数と負例での出現回数の組の頻度情報を示している．生成した負例相互作用データにおける各エンハンサーと各プロモーターの出現回数が正例と同程度になっていることが分かる．

さらに，図 4 は，細胞株 K562 に対する，3 種類の正負例データセット (TargetFinder, EP2vec, 我々の負例データ) の EP2vec モデルでの交差検証の予測精度 (F 値) の比較結果を示している．TargetFinder や EP2vec の正負例データセットでは，ランダムに訓練用とテスト用に分割する交差検証に比べ染色体単位で訓練用とテスト用に分割する交差検証では予測精度が大きく低下している．一方，本手法の負例を用いた場合は，染色体単位の交差検証での予測精度の低下は生じてない．

5. 考察

エンハンサー・プロモーター間相互作用予測問題に関して，正例データから構成する有向グラフに対し最大フロー問題を解くことで各エンハンサーと各プロモーターの出現回数が正例と負例で同程度となるような負例データの構成に成功している．この正負例データセットを用いた EP2vec

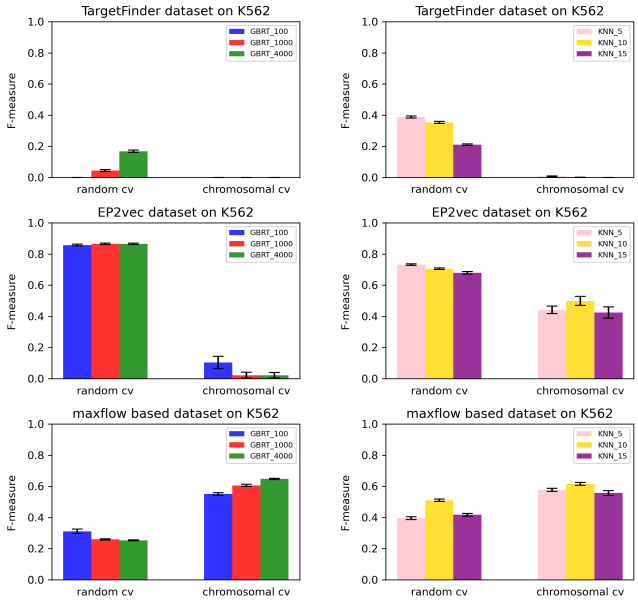


図 4: 細胞株を K562 とした時の各正負例データセットを用いて交差検証を行った EP2vec 予測モデルでの F 値．上段と中段と下段でそれぞれ TargetFinder と EP2vec と本研究で構成した正負例データセットでの F 値を示す．データセットをランダムに訓練用とテスト用に分割する一般的な交差検証（random cv）と染色体単位で訓練用とテスト用に分割する交差検証（chromosomal cv）の 2 通りの交差検証での F 値を示す．

予測モデルの染色体単位の交差検証では，TargetFinder や EP2vec の正負例データセットに見られた予測精度の低下が生じてないので，エンハンサー・プロモーター間相互作用に関する本質的な特徴を学習していることが示唆される．

また，本研究の対象はエンハンサー・プロモーター間相互作用であったが，エンハンサー・プロモーターに限らず様々な 2 つのオブジェクト間の相互作用予測問題の負例生成に応用可能と思われる．

参考文献

[1] Smemo, S., Campos, L. C., Moskowitz, I. P., Krieger, J. E., Pereira, A. C. and Nobrega, M. A.: Regulatory Variation in a TBX5 Enhancer Leads to Isolated Congenital Heart Disease, *Human Molecular Genetics*, Vol. 21, No. 14, pp. 3255–3263 (2012).

[2] Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., Valen, E., Li, K., Schwarzfischer, L., Glatz, D., Raithel, J., Lilje, B., Rapin, N., Bagger, F. O., Jørgensen, M., Andersen, P. R., Bertin, N., Rackham, O., Burroughs, A. M., Baillie, J. K., Ishizu, Y., Shimizu, Y., Furuhata, E., Maeda, S., Negishi, Y., Mungall, C. J., Meehan, T. F., Lassmann, T., Itoh, M., Kawaji, H., Kondo, N., Kawai, J., Lennartsson, A., Daub, C. O., Heutink, P., Hume, D. A., Jensen, T. H., Suzuki, H., Hayashizaki, Y., Müller, F., Forrest, A. R. R., Carninci, P., Rehli, M. and Sandelin, A.: An Atlas of Active Enhancers across Human Cell Types and Tissues, *Nature*, Vol. 507, No. 7493, pp. 455–461 (2014).

[3] Jiang, R.: Walking on Multiple Disease-Gene Networks to Prioritize Candidate Genes, *Journal of Molecular Cell Biology*, Vol. 7, No. 3, pp. 214–230 (2015).

[4] Whalen, S., Truty, R. M. and Pollard, K. S.: Enhancer-Promoter Interactions Are Encoded by Complex Genomic Signatures on Looping Chromatin, *Nature genetics*, Vol. 48, No. 5, pp. 488–496 (2016).

[5] Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S. and Dekker, J.: Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome, *Science (New York, N.Y.)*, Vol. 326, No. 5950, pp. 289–293 (2009).

[6] Zeng, W., Wu, M. and Jiang, R.: Prediction of Enhancer-Promoter Interactions via Natural Language Processing, *BMC Genomics*, Vol. 19, No. 2, p. 84 (2018).

[7] Singh, S., Yang, Y., Póczos, B. and Ma, J.: Predicting Enhancer-Promoter Interaction from Genomic Sequence with Deep Neural Networks, *Quantitative biology (Beijing, China)*, Vol. 7, No. 2, pp. 122–137 (2019).

[8] Hong, Z., Zeng, X., Wei, L. and Liu, X.: Identifying Enhancer–Promoter Interactions with Neural Network Based on Pre-Trained DNA Vectors and Attention Mechanism, *Bioinformatics*, Vol. 36, No. 4, pp. 1037–1043 (2月 15, 2020).

[9] Liu, S., Xu, X., Yang, Z., Zhao, X., Liu, S. and Zhang, W.: EPIHC: Improving Enhancer-Promoter Interaction Prediction by Using Hybrid Features and Communicative Learning, *IEEE/ACM transactions on computational biology and bioinformatics*, Vol. PP (2021).

[10] Xi, W. and Beer, M. A.: Local Epigenomic State Cannot Discriminate Interacting and Non-Interacting Enhancer–Promoter Pairs with High Accuracy, *PLoS Computational Biology*, Vol. 14, No. 12, p. e1006625 (2018).

[11] Cao, F. and Fullwood, M. J.: Inflated Performance Measures in Enhancer–Promoter Interaction-Prediction Methods, *Nature Genetics*, Vol. 51, No. 8, pp. 1196–1198 (2019).

[12] Thomas H. Cormen, Charles E. Leiserson, R. L. R. and Stein, C.: *Introduction to Algorithms Third Edition*, The MIT Press.

[13] Le, Q. V. and Mikolov, T.: Distributed Representations of Sentences and Documents, *arXiv:1405.4053 [cs]* (2014).