

ncRNA-疾患関連性の予測法の改良

母里 卓也^{†1} 林田 守広^{†2} 阿久津達也^{†3} ホセ・ナチエル^{†4}

近年、non-codingRNA (ncRNA) の突然変異・制御異常とヒトの疾患の関係についていくつもの報告がなされている。したがって、ncRNA-疾患の関連性予測はバイオインフォマティクスにおいて重要な研究課題となっている。我々は以前の研究において、生物学的配列と、ncRNA-標的と標的-疾患との関連から定義された三層ネットワークを使用する方法を示した。本研究では、予測精度を向上させるために、最適化プロセスおよび生物学的配列間の類似性尺度についての詳細な解析を含むいくつかの方法を調べた。これらの計算法、および、その予測結果について報告する。

Improvements on ncRNA-disease association predictions

Takuya Mori^{†1} Morihiro Hayashida^{†2} Tatsuya Akutsu^{†3} Jose C. Nacher^{†4}

Recent research has shown multiple evidences on the relation between mutations and deregulations of non-coding RNAs and specific human diseases. Therefore, the development of computational models that predict ncRNA-disease associations has become an important research topic in bioinformatics. Here, we report on a proposed method of ncRNA-disease association prediction based on sequence and tripartite network, which is defined using intermediate associations between ncRNAs-targets and targets-diseases. Moreover, in order to enhance model predictions, several methods were examined, including optimization processes and a more detailed analysis of the similarity between sequences. This extended abstract briefly presents and discusses the results from the computational method together with those from the improved analysis.

1. はじめに

近年の研究では、non-coding RNA (ncRNA) の生物学的機能、トランスクリプトーム、および幅広い生物におけるあらゆるサイズの ncRNA の調節機能が研究されている[1]。siRNA (small interfering RNA)、miRNA (micro RNA) および piRNA (Piwi-interacting RNA) は、small non-coding RNA の主なタイプである (30 塩基未満)。200 塩基以上の long non-coding RNA (lncRNA) は、クロマチン修飾や転写制御などの遺伝子調節の基本的プロセスに関連している[2]。計算機実験および生物学的実験により、短い ncRNA の改変、脱調節白血病、肝細胞および結腸癌ならびに神経変性障害、心臓血管疾患および免疫介在性疾患のような種々のタイプの癌を引き起こす可能性があることが知られている[3,4]。したがって、ncRNA とヒト疾患との間の関連を予測する計算方法を開発することが重要である。近年、三者間ネットワーク構造と資源配分法やシーケンス情報を組み合わせた方法が提案されている[5,6]。

2. 研究方法

我々は以前に、ncRNA-標的-疾患三者ネットワークにおけるリソース割り当て法を、生物学的配列から得られた類似情報と統合する方法を提案した。生体情報と組み合わせ

れたこの方法は、ネットワーク全体に伝達され、最終的に新しい関連を推定するために利用される[6]。この方法は、既存の手法である ncPred アルゴリズム[5]の結果を上回っていた。しかし、ncRNA と疾患との間の関連性を予測できない部分に関して改善の余地がある。そこで、2 つの代替方法を検討した。

まず、[5]で使われた RBF カーネルを文字列カーネルに置き換えることを考えた。Smith-Waterman (SW) アルゴリズムは、2 つの配列間の類似性を局所ギャップ配列によって評価する。しかし、SW アルゴリズムのスコアはカーネルの定義を満たさない。そこで、 $K_{LA}^{(\beta)}$ と呼ばれる局所アライメント・カーネルが以前の研究[7]で開発され、私たちはそのカーネルを実装した。 $K_{LA}^{(\beta)}$ は、 β が無限大になるとき SW アルゴリズムスコアに収束する。我々の研究では、 β を 1 に設定し、ギャップ開始および延長コストを、それぞれ $d = -11$ および $e = -1$ に設定した。ROC 解析により、Helwak データセットの曲線下面積(AUC)を求めたとき、カットリンク無しの値は 0.69、カットリンクの有り値では 0.59 が得られた。AUC の値は有用であるといえるが、RBF カーネルを使用した予測 (表 1 を参照) よりも少し低くなっている。二つ目の方法では、非負行列因子分解データ統合方法[9,10]の変形版を検討した。入力行列 R_{ij} は、行と列が ncRNA、ターゲット (タンパク質)、および疾患 (図 1 参照) などのデータセットでありブロック行列に関係する行列を表している。入力行列 R_{ij} は、 r 個のデータ型の利用可能なすべての関連を含むブロック行列 R に結合される。

^{†1} 東邦大学大学院 理学研究科 情報科学専攻
Graduate School of Science, Toho University

^{†2} 松江工業高等専門学校 電気情報工学科
Department of Electrical Engineering and Computer Science, National Institute of Technology, Matsue College

^{†3} 京都大学 化学研究所 バイオインフォマティクスセンター
Bioinformatics Center, Institute for Chemical Research, Kyoto University

^{†4} 東邦大学 理学部 情報科学科
Department of Information Science, Toho University

一方、 Θ_{ij} 行列は因子分解アルゴリズムのデータ制約を考慮し、同じ種類の要素間の関連性を表現する(ncRNA-ncRNA 関係や標的- 標的関係など) (図 1 参照)。

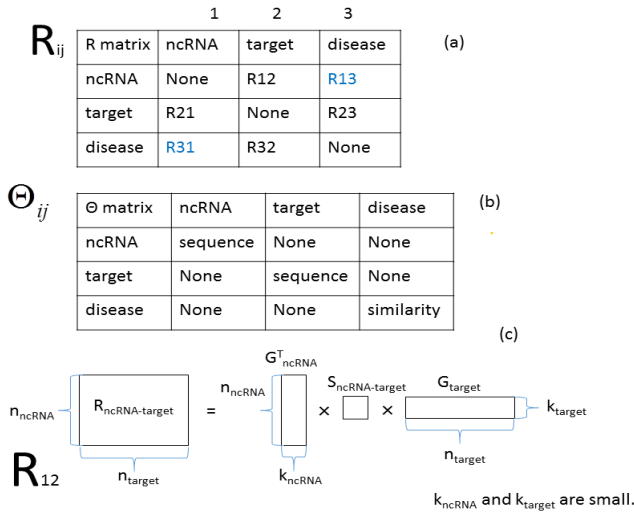


図 1 行列 R_{ij} と Θ_{ij} の概要

この方法は、 R_{ij} を非負行列に三因子分解することにある。

$$R_{ij} \approx G_i S_{ij} G_j \quad (1)$$

G 行列、 S 行列の値は、初期化のためランダムに $[0,1]$ 一様分布からサンプリングされる。次に、目的関数 J は、以下のように定義され最小化される。

$$\min J(G; S)_{G \geq 0} = \sum_{R_{ij}} \|R_{ij} - G_i S_{ij} G_j\|^2 + \sum_{f=1}^F \text{tr}(G^T \theta^{(f)} G) \quad (2)$$

関数の最初の項は元の行列と三元化された行列の差のフロベニウスノルムである。2 番目の項は制約行列に依存している。今回、 $F=1$ であるのは、配列類似性を 1 種類しか使用しなかったためである。アルゴリズムの反復の停止基準を $T=10^{-4}$ に設定し、1 回の計算で 10 回実行した。倍率 λ は 200 に設定した。標準的な方法は入力データとして R_{13} の情報を使用するが (図 1 参照)、入力データとして R_{13} を必要としない ($R_{13}=0$) 別の方法を検討した。この方法の詳細は、口頭で発表する。また、本研究では疾患類似性は含めていない ($\Theta_{33}=0$) が、含めることは可能である。

3. 結果とまとめ

ここでは、両方の検討手法の結果について簡単に報告する。まず、SW アルゴリズムを元にしたカーネルから類似情報を得て、ncRNA-ターゲット-疾患の三者ネットワークを用いたリソース割当手法の結果について議論する。ROC 分析により、Helwak データセットに対する AUC スコアを求めたところ、カットリンク無しの値は 0.69、カットリンク有りの値では 0.59 が得られた。AUC の値は有用で

あるといえるが、RBF カーネルを使用した予測 (表 1 を参照) よりもまだ少し低くなっている。

次に、入力情報を設定しない場合の改良因子分解法の予測結果を表 1 に示す。Chen データセット[11]の場合、AUC スコアはそれぞれリンクカットなしとリンクカットありで

		Alaimo	our previous work	Kernel LA	Matrix factorization
Helwak	cut link	0.502	0.774	0.590	0.777
	without	0.673	0.854	0.694	0.989
Chen	cut link	0.502	0.570		0.632
	without	0.730	0.750		0.855

表 1 各手法による AUC の結果

0.85 と 0.63 である。Helwak データセットの場合、AUC スコアはそれぞれリンクカットなしとリンクカットありで 0.98 および 0.77 である。

結果は、 R_{13} に関する情報を必要としない改善された因子分解法により、ncRNA-疾患の予測を改善する可能性があることを示している。また新たな ncRNA-疾患関連を同定し、それらが他の独立の生物学的実験研究においても示されているかどうかを評価するために、さらなる分析が必要である。

参考文献

- [1] Mercer TR, Dinger ME, Mattick JS. Long non-coding RNAs: insights into functions. Nat Rev Genet. 2009;10:155–9. 2.
- [2] Wang KC, Chang HY. Molecular mechanisms of long noncoding RNAs. Mol Cell. 2011;43(6):904–14. 3.
- [3] Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. 2012;22(9):1775–89. 4.
- [4] Wapinski O, Chang HY. Long noncoding RNAs and human disease. Trends Cell Biol. 2011;21(6):354–61.
- [5] Alaimo S, Giugno R, Pulvirenti A. ncPred: ncRNA-disease association prediction through tripartite network-based inference. Front Bioeng Biotechnol. 2014;2:71
- [6] Mori et al, ncRNA-disease association prediction based on sequence information and tripartite network, BMC Systems Biology 2018;12 (Suppl 1):37
- [7] Saigo, H, Vert, J-P., Ueda, N. and Akutsu, T. Protein homology detection using string alignment kernels, Bioinformatics 2004, 1682-169.
- [8] Helwak A, Kudla G, Dudnakova T, Tollervey D. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. Cell. 2013; 153(3):654–65.
- [9] Zitnik, M and Zupan, B. Data Fusion by matrix factorization. IEEE Trans. Pattern. Anal. Mach. Intell. 37, 41-53 (2015).
- [10] Marini, S, Vitali, F, Rampazzi, S, Demartini, A. and Akutsu, T. Protease target prediction via matrix factorization <http://dx.doi.org/10.1101/275024>.
- [11] Chen G, Wang Z, Wang D, Qiu C, Liu M, Chen X, et al. LncRNADisease: a database for long-non-coding RNA-associated diseases. Nucleic Acids Res. 2013;41:D983–6. 16.