社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－MPS－67 (30)
2007－BIO－11 (30)
2007／12／21

整数計画法に基づくRNA間相互作用予測

市原 慶太郎† 加藤 有己‡ 渡邊 芳英*

†同志社大学大学院 工学研究科
‡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
*同志社大学 工学部 電子工学科

概要 遺伝子発現の転写後調節に関わるRNA間の相互作用が注目されている．一方，最適化ソルバーの能力が急速に高性能化し，いままでは計算が困難な問題であった規模の大きな整数計画問題も比較的高速で解けるようになってきている．したがって，このようなRNA相互作用による2次構造予測問題を整数計画問題として定式化することができれば，最適化ソルバーを用いて解くことが実現可能な時代となってきた．本研究では，2つのRNAの相互作用による結合2次構造を，整数計画法を用いて予測する方法を提案する．また，提案手法による結合2次構造予測のいくつかの実験結果を示す．

RNA-RNA Interaction Prediction Based on Integer Programming

Keitaro Ichihara†, Yuki Kato‡ and Yoshihide Watanabe*

†Graduate School of Engineering, Doshisha University
‡Bioinformatics Center, Institute for Chemical Research, Kyoto University
*Department of Electronics, Faculty of Engineering, Doshisha University

Abstract RNA-RNA interaction involved in post-transcription regulation of gene expression has received much attention. On the other hand, the performance of optimization solvers has made dramatic progress in the sense that they have been able to solve large-scale optimization problem in a relatively short time. It might be a good strategy to focus on the formulation of the problem to be solved as a mathematical programming problem and leave it to a solver. We propose an algorithm for predicting the joint secondary structure of two interacting RNAs by solving certain integer programming problem. In addition, we show some results on computational experiments on the prediction of the joint secondary structure by using our proposed method.

1 はじめに

実験によって，すべての生命情報を解明することは不可能である．そのため，情報技術による生命情報の解析と理解を目的としたバイオインフォマティクスが重要視されている．バイオインフォマティクスでは核酸やタンパク質配列の相同性検索，構造，機能の予測などが行われる．本論文ではRNAの2次構造予測に焦点をあてる．RNAやタンパク質の多くは固有の立体構造を持つ．配列データのみから，その立体構造や立体構造の骨組みとなる2次構造を推定しようというのが構造予測である．このような構造予測のために，アルゴリズム理論や理論計算機科学といった分野における様々な数理的な手法が駆使されている．このような手法を使うためにまず重要なことは対象とする問題を数理的に定式化することである．

近年，遺伝子発現を抑制するRNA干渉(RNAi)と同様に，それに似た働きを持つアンチセ

ンスRNAの調節に対する多くの研究がなされている．アンチセンスRNAは，mRNAと結合することで，遺伝子発現の転写後の調節に関与している．そのような2つの相互作用を持つRNAの塩基対構造は，結合2次構造と呼ばれている．結合2次構造の中には多くのループ間相互作用が観察され，そのなかにはキッシングヘアピンループと呼ばれるものが含まれている．このような結合2次構造を予測するためのひとつの方法は，予測問題を水素結合の自由エネルギーを最小化する最適化問題と考えることである．そのような最適化問題を解くために，動的計画法アルゴリズム（DP)を用いたいくつかの手法が開発されてきた[1, 2, 6]．しかし我々が知る限りでは，一般的な整数計画（IP)問題を扱ってこのような問題を扱った研究は存在しない．そのような理由から我々はこのような結合2次構造の予測問題を，0-1整数計画問題として定式化してそれを解くという手法を提案する．整数計画問題として定式化することの長所として問題に対する融通性や，設計の自由度の高さが上げられる．すなわち，類似問題に対しての定式化の変更が容易であり，与えられた問題の制約条件の変化に柔軟に対応することができる．短所として計算量が非常に多いため，サイズが大きい問題を解くことが困難でああることがあげられる．しかし近年，コンピュータのハードウェアの急速な進歩とともに，商用の最適化ソルバーの能力が劇的に進化し，いままで解くことが困難であった大規模整数計画問題も妥当な計算時間で解くことができるようになってきている．

すでに述べたように，2つのRNA分子間での結合2次構造を予測する問題を，相互作用を表す塩基対の自由エネルギーを最小化する問題として取り扱うことができる．本稿では，この問題を塩基対の重みつきの個数を最大化する整数計画問題として定式化し，それを最適化ソルバーを用いて解くという手法を提案する．われわれの提案手法の有効性の検証を行うための計算機実験に用いる最適化ソルバーとして，様々な分野で広く使われ，その性能について定評があるソルバーのひとつであるILOG社のCPLEXを採用した．

2 RNA間相互作用予測問題

定義 1. (RNA結合2次構造)

$\Sigma = \{\mathrm{A,C,G,U}\}$ とし，Σの要素からなる任意の長さのの塩基配列全体をΣ^*と表す．また$a \in \Sigma$に対して，$|a|$は配列の長さを表すものとする．2つの塩基配列$a = a_1 \cdots a_n \in \Sigma^*(|a| = n$，5'-3'方向)，$b = b_1 = \cdots b_m \in \Sigma^*$ ($|b| = m$，3'-5'方向)をとる．そのとき，αはaまたはbを表すものとして(すなわち$\alpha \in \{a, b\}$として)，以下の2つの条件を満たす塩基の位置の対(i, j)からなる集合をS_αで定義する：

- $1 \leq i < i+1 < j \leq |\alpha|$,
- $\forall(i,j), \forall(i',j') \in S_\alpha; i = i' \Leftrightarrow j = j'$.

最初の条件は，各塩基配列において，隣り合った塩基は水素結合対にならないということを表し，2番目の条件は，各塩基配列において，各塩基が高々1個の塩基とのみ水素結合対を作ることを表す．

次に，a-b間で以下の条件を満たす塩基の位置の対(k, l)からなる集合をS_{ab}とする：

- $(\exists i; (i,k) \in S_a$または$(k,j) \in S_a)$，かつ$(\exists j; (j,l) \in S_b$または$(l,j) \in S_b) \Rightarrow (k,l) \notin S_{ab}$,

- $\forall(k,l),(k',l') \in S_{ab}; k = k' \Leftrightarrow l = l'$.

最初の条件は，S_a, S_b のいずれか一方の塩基配列において，水素結合対として現れる塩基，すわなち S_a または S_b に含まれる水素結合対に現れる塩基は，a, b 間の結合対には現れないことを表している．2番目の条件は a, b の塩基は S_{ab} の水素結合対において高々1度しか使われないことを表している．このとき，$S = (S_a, S_b, S_{ab})$ を a と b の結合2次構造と呼ぶ．

$S_\alpha(\alpha \in \{a, b\})$ に属する各塩基対は内部結合しているといい，S_{ab} に属する各塩基対は外部結合しているという．RNA は内部結合および外部結合をすることにより自由エネルギーが小さくなり安定化する．したがって，RNA2 次構造予測問題は，自由エネルギー最小化問題として定式化することができる．ここでは，水素結合を行う塩基対として Watson-Crick 対 (A-U,C-G の対) のみを考え，結合2次構造を予測する問題を，塩基対の個数を2つの Watson-Crick 対に適当な重みを与えた上で最大化するという組み合わせ最適化問題（整数計画問題）として定式化する．

定義 2. (RNA 間相互作用予測問題)

入力：2つの RNA 配列 $a = a_1 \cdots a_n, b = b_1 \cdots b_m$.
出力：以下のスコアを最大にする a と b の結合2次構造 $S = (S_a, S_b, S_{ab})$：

$$\sum_{(i,j)\in S_a} w(a_i, a_j) + \sum_{(i,j)\in S_b} w(b_i, b_j) + \sum_{(k,l)\in S_{ab}} w(a_k, b_l).$$

ここで，スコア関数 w は次のように指定される：

$$w(\xi_1, \xi_2) = \begin{cases} \{\xi_1, \xi_2\} = \{\mathrm{A}, \mathrm{U}\} \text{ のとき } 2,\ \{\xi_1, \xi_2\} = \{\mathrm{C}, \mathrm{G}\} \text{ のとき } 3. \\ \text{上記以外のとき}, -\infty. \end{cases}$$

文献 [1, 4, 6] では，RNA 結合2次構造 $S = (S_a, S_b, S_{ab})$ において，さらに以下の制約を考えている．

条件 (Pseudoknot free)

1. S は**内部シュードノット**を含まない．すなわち，
   $\forall(i,j),(i',j') \in S_\alpha(i < i', \alpha \in \{a,b\}) \Rightarrow i < i' < j < j'$ が成り立たない．
2. S は**外部シュードノット**を含まない．すなわち，
   $\forall(k,l),(k',l') \in S_{ab}(k < k') \Rightarrow l' < l$ が成り立たない．

条件 1 は $S_\alpha(\alpha \in \{a,b\})$ の塩基対が内部結合しているときに交差することがないことを意味し，条件 2 は S_{ab} の塩基対が外部結合しているときに交差することがないということを意味している．

本研究においても，このシュードノットなし制約条件を採用し，その制約のもとで結合2次構造の予測問題を考察する．その理由は，現在までに見つかっている RNA 結合2次構造は，内部，外部両シュードノットを含まないと報告されているからである [1]．このような制約条件をつけることにより探査すべき範囲が狭くなり整数計画問題の変数を減らすことができるので，計算量を減らすことができる．良く知られているように，単独の RNA2 次構造予測問題を自由エネルギーを最小化する最適化問題として捉えたものは，シュードノットなしという制約条件のもとで動的計画法に基づく Nussinov アルゴリズム [7] によって解かれているが，その類似をたどることにより結合2次構造予測問題についても動的計画法に基づき $O(n^3m^3)$ 時間で予測するアルゴリズムが設計されている [1, 4, 6].

3 整数計画問題への定式化

本論文では，シュードノットなしの制約条件を満たすRNA結合2次構造のクラスを，整数計画法に基づいて予測する手法を提案する．整数計画問題に用いる変数を以下のように設定する:

$$x_{ij} = \begin{cases} 1 & (a_i \text{ と } a_j \text{ が塩基対を形成する}), \\ 0 & (\text{それ以外}). \end{cases}$$

$$y_{ij} = \begin{cases} 1 & (b_i \text{ と } b_j \text{ が塩基対を形成する}), \\ 0 & (\text{それ以外}). \end{cases}$$

$$z_{kl} = \begin{cases} 1 & (a_k \text{ と } a_l \text{ が塩基対を形成する}), \\ 0 & (\text{それ以外}). \end{cases}$$

ここで，x 変数については $1 < i < j < n$ であり，y 変数については $1 < i < j < m$ であるとする．これらの変数は，塩基対があるときには1になり，塩基対がないときには0をとるという2つの値だけをとる変数であり，考える整数計画問題は0-1整数計画問題となる．ただ，このままでは，塩基対の長さが大きくなると変数の数が非常に大きくなり，整数計画問題を解くことが困難になる．そこでやや作為的ではあるが変数の数を減らすために，自然数パラメータ L を導入し，x, y 変数の添え字のとる値の範囲を狭める．その結果として各変数の添字のとる範囲は以下のようになる：

- $x_{ij} (i = 1, \ldots, n - L; j = i + L, \ldots, n)$,
- $y_{ij} (i = 1, \ldots, m - L; j = i + L, \ldots, m)$,
- $z_{kl} (k = 1, \ldots, n; l = 1, \ldots, m)$.

この最初の2つの条件は，各塩基配列 a, b の内部においては長さが L 以上離れた塩基とのみ塩基対を作るということを意味する．RNA2次構造においては，塩基対に挟まれた連続領域をループと呼ぶ．したがって，L はRNA分子がとる2次構造のループの長さを反映する定数であり，このパラメータ L はループ長と呼ばれる．ループ長の設定について特にルールがあるわけではないが今回の計算機実験では $L = 7$ を採用している．最大化すべき目的関数を設定するため，以下のスコアを与える：

- 配列 a に対し，$w_{ij}^{(a)}$ は $\{a_i, a_j\} = \{A, U\}$ のとき2，$\{a_i, a_j\} = \{C, G\}$ のとき3である．
- 配列 b に対し，$w_{ij}^{(b)}$ は $\{b_i, b_j\} = \{A, U\}$ のとき2，$\{b_i, b_j\} = \{C, G\}$ のとき3である．
- a-b 間に対し，$w_{kl}^{(ab)}$ は $\{a_k, b_l\} = \{A, U\}$ のとき $2 * 0.8$，$\{a_k, b_l\} = \{C, G\}$ のとき $3 * 0.8$ である．

このスコアは内部の塩基対の水素結合についてはすでに与えた値を用いているが，外部の水素結合についてはその値に0.8をかけた値になっている．以上で定義された変数と目的関数を用いて，RNA結合2次構造予測問題は，次のような0-1整数計画問題として定式化される．その際の制約条件は以下のようになる：

$$\text{maximize} \quad \sum_{i=1}^{n-L} \sum_{j=i+L}^{n} w_{ij}^{(a)} x_{ij} + \sum_{i=1}^{m-L} \sum_{j=i+L}^{m} w_{ij}^{(b)} y_{ij} + \sum_{k=1}^{n} \sum_{l=1}^{m} w_{kl}^{(ab)} z_{kl}$$

$$\text{subject to} \quad \sum_{j=i+L}^{n} x_{ij} \leq 1 \quad (i = 1, \ldots, L), \tag{1}$$

$$\sum_{h=1}^{i-L} x_{hi} + \sum_{j=i+L}^{n} x_{ij} \leq 1 \quad (i = L+1, \ldots, n-L), \tag{2}$$

$$\sum_{i=1}^{j-L} x_{ij} \leq 1 \quad (j = n-L+1, \ldots, n), \tag{3}$$

$$x_{ij} + x_{kl} \leq 1 \quad (1 \leq i < k < j < l \leq n), \tag{4}$$

$$\sum_{j=i+L}^{m} y_{ij} \leq 1 \quad (i = 1, \ldots, L), \tag{5}$$

$$\sum_{h=1}^{i-L} y_{hi} + \sum_{j=i+L}^{m} y_{ij} \leq 1 \quad (i = L+1, \ldots, m-L), \tag{6}$$

$$\sum_{i=1}^{j-L} y_{ij} \leq 1 \quad (j = m-L+1, \ldots, m), \tag{7}$$

$$y_{ij} + y_{kl} \leq 1 \quad (1 \leq i < k < j < l \leq m), \tag{8}$$

$$\sum_{l=1}^{m} z_{kl} \leq 1 \quad (k = 1, \ldots, n), \tag{9}$$

$$\sum_{k=1}^{n} z_{kl} \leq 1 \quad (l = 1, \ldots, m), \tag{10}$$

$$z_{ij} + z_{kl} \leq 1 \quad (1 \leq i < k \leq n, 1 \leq l < j \leq m), \tag{11}$$

$$2x_{kj} + \sum_{l=1}^{m} z_{kl} + \sum_{l=1}^{m} z_{jl} \leq 2 \quad (k = 1, \ldots, n-L; j = k+L, \ldots, n), \tag{12}$$

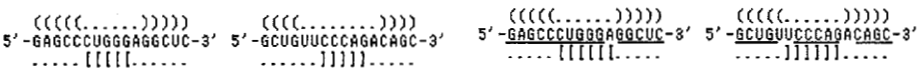
$$2y_{lj} + \sum_{k=1}^{n} z_{kl} + \sum_{k=1}^{n} z_{kj} \leq 2 \quad (l = 1, \ldots, m-L; j = l+L, \ldots, m), \tag{13}$$

$$x_{ij} \in \{0,1\} \quad (i = 1, \ldots, n-L; j = i+L, \ldots, n)$$

$$y_{ij} \in \{0,1\} \quad (i = 1, \ldots, m-L; j = i+L, \ldots, m)$$

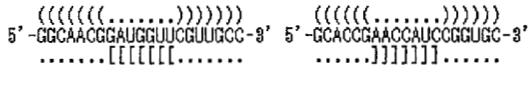
$$x_{kl} \in \{0,1\} \quad (k = 1, \ldots, n; L = 1, \ldots, m)$$

制約条件 (1)〜(3) は，a の各塩基が高々ひとつの内部結合対にしか含まれないことを示している．条件が 3 つにわかれたのは，ループ長パラメータ L があるためであり，i が L 以下なら i より前の塩基と対をつくることができず，j が $n-L$ より大きいときは j より後の塩基と対をつくることができない．それ以外の場合は i の前，または後ろに対となる塩基が存在する可能性があるが，その場合は前か後ろの高々一方の塩基としか対をつくることはできない．それらの条件がそれぞれ (1)，(3) および (2) で表されている．(4) と (8) は内部シュードノットを禁止する条件である．(5)〜(7) は塩基配列 a の内部結合対に対する制約条件 (1)〜(3) を塩基配列 b に対する同様の制約条件に置き換えたものである．条件 (9)，(10) は塩基配列 a の塩基と塩基配列 b の塩基がそれぞれ高々1 つの外部結合対にしか含まれることはないという条件である．条件 (11) は外部シュードノットを禁ずる条件である．条件 (12) と (13) はそれぞれの塩基対に含まれる塩基が，内部結合対と外部結合対に同時に含まれることはないことを示す．

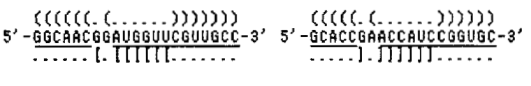


(a) Known joint secondary structure (b) Predicted joint secondary structure

図 1: Tar-Tar*の結合 2 次構造



(a) Known joint secondary structure



(b) Predicted joint secondary structure

図 2: RNAIi-RNAIIi の結合 2 次構造

4 実験結果

本論文で提案した，整数計画問題として定式化した RNA 間相互作用予測法の有効性を検証するため，いくつかの比較的簡単な例で計算機実験を行った．ここで，整数計画問題を解くために，商用のソルバーとして定評のある ILOG 社の CPLEX[8] を用いた．計算機実験における計算機環境は Intel Core2 CPU T7600 2.33GHz，2.33GHz，3.00GB RAM，OS は Windows XP である．ソルバー CPLEX を用いて整数計画問題を解くことにより RNA 結合 2 次構造が予測できた RNA 配列の組は Tar-Tar*[3] および RNAIi-RNAIIi[5] であり，その長さは 16-16，21-19 である．これらの RNA 配列は相互作用をすることが確かめられている RNA 配列の例のなかから選んだ．図 1，図 2 において，内部結合は丸括弧で，外部結合は角括弧で表されている．予測精度を評価する尺度として，感度と特異度を計算した．感度は，実データにおいて実際に存在する結合対の個数に対する，アルゴリズムによって予測された結合対のうち実際に存在する結合対 (正しい結合対) の数の割合で定義される．また，特異度はアルゴリズムによって予測された結合対全体の個数に対する，正しく予想された結合対の個数の割合として定義される．予測構造を図 1(b)，図 2(b) に示す．図において，正解と一致した予測塩基対には下線が引かれている．また，予測精度を表 1 に示す．

5 おわりに

計算機実験の結果では，16-16，21-19 という比較的短い RNA 配列に対して，本稿で提案した整数計画問題として定式化してそれを解くという予測法がかなりよい精度をもつことを確かめることができた．しかし，より長い配列に対して計算機実験を行ったところメモリーが不足し計算実行が不可能となった．これは当然予想されたことであり，最新の高速ソルバーをもってしても大規模な整数計画問題を解くことは難しいということを実感せざるを得ない結論となった．このような計算量の大きくなるという問題をある程度克服するには（可能なら）ループ長 L 定数を大きくとるなどして探索の範囲を狭めて，変数の数を減らすことが考えられるが，それ

表 1: 予測精度

RNA 複合体	Tar-Tar*	RNAIi-RNAIIi
感度 [%]	100.0	85.0
特異度 [%]	87.5	85.0

も一定の限界があると考えられる．このような問題を整数計画問題として定式化する大きな利点は，問題の変更に対する柔軟性にある．すなわち，少しモデルを変更して，制約条件を変更することが比較的簡単に出来ることである．したがって，モデルまたは同じモデルのパラメータを少しずつ変更しながら実験を行うことが比較的容易に実行できる．例えば，ループ長を変更する，コストを変更する，シュードノットなしの条件からシュードノットありの条件に変更するなどは容易にできる．このような観点にたてば，本研究は RNA 結合 2 次構造予測の方法についてもっともよい数学的なモデルはどのようなものであるか探るための第 1 歩であるとみることもできるであろう．

参考文献

[1] Alkan, C., Karakoc, E., Nadeau, J. H., Sahinalp, S. C. and Zhang, K.: RNA-RNA Interaction Prediction and Antisense RNA Target Seach, *Journal of Computational Biology*, Vol. 13, pp. 267–282 (2006).

[2] Andronescu, M., Zhang, Z. C. and Condon, A.: Secondary Structure Prediction of Interaction RNA Molecles, *Journal of Molecular Biology*, Vol. 345, pp. 987–1001 (2005).

[3] Chang, K. Y. and Tinoco Jr, I.: The Structure of an RNA "Kissing" Hairpin Complex of the HIV TAR Hairpin Loop and Its Complement, *Journal of Molecular Biology*, Vol. 269, pp. 52–66 (1997).

[4] Kato, Y., Akutsu, T. and Seki, H.: A Grammatical Approach to RNA-RNA Interaction Prediction, *AIP Conference Proceedings*, Vol. 952, 2007 International Symposium on Computational Models for Life Sciences (CMLS'07), pp. 197–206 (2007).

[5] Lee, A. J. and Crothers, D. M.: The Solution Structure of an RNA Loop-Loop Complex: the ColE1 Inverted Loop Sequence, *Structure*, Vol. 15, No. 6, pp. 993–1005 (1998).

[6] Pervouchine, D. D.: IRIS: Intermolecular RNA Interaction Search, *Proceedings of the 15th International Conference on Genome Informatics (GIW2004)*, pp. 92–101 (2004).

[7] Nussinov, R., Pieczenik, G., Griggs, J. R. and Kleitman, D.J.: Algorithms for Loop Matchings, *SIAM Journal of Applied Mathematics*, Vol. 35, No. 1, pp. 68–82 (1978).

[8] http://www.ilog.com/products/cplex