Vol.37 No.10 情 報 処 理 Oct. 1996

解 説

ゲノム情報

8. 超並列計算機を用いたRNA2次構造の予測と視覚化†

中谷明弘†† 山本健二††† 米澤明憲††††

1. はじめに

RNAはその化学的側面を単純化すると，4種類の塩基（Adenine, Uracil, Guanine, Cytosine）が一本鎖状に連なったものと考えられる．この巨大分子はA-U, G-C, G-U 間に水素結合を生じ，エネルギー的に安定な高次構造を作ることになる．RNAの機能を決定していると考えられるこの構造を予測することは，これらRNAが引き起こすさまざまな現象，たとえば，動植物にとりつくRNAウィルスの解析などに貢献するものと期待されている．3次元空間中での構造予測は困難なため，RNAが2次元平面上に存在すると制限した2次構造予測が広く行われている．以下では，RNA2次構造予測とその結果の視覚化を超並列計算機を用いて行った例を紹介する．

2. RNAの2次構造予測

従来から多くの2次構造予測法[1]が試みられているが，その多くは最適解を1つ求めるものである．ところが，図-1(a) にあげた単純な塩基列が，ほとんど同エネルギーの2次構造を図-1(b)のように複数生じている例が示す通り，最適解が1つ求められても必ずしも十分ではないことがわかる．実際に，ある種のRNAは「槍状」と「放射状」の2次構造を持ち，それぞれが異なった局面で機能を発揮していることが知られている．このため，最安定解のみではなく，最安定解から ΔKkcal（ΔKはパラメータ）だけ不安定な準安定解もすべて求める方法[2)~4)]が必要とされる．

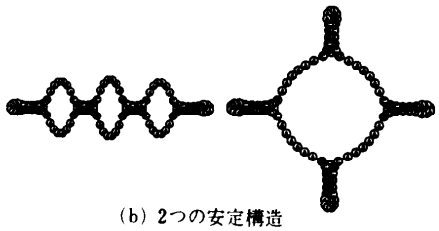
2次構造は図-2のような連続した水素結合（スタック領域）の組合せからなる．図中，円が塩基を表し，点線が塩基間の水素結合を表す．塩基は5'端を1として番号がつけられ，塩基列中の位置が表される．スタック領域の塩基列中での位置もスタック領域を構成する水素結合の塩基番号によって表される．この塩基番号の大小を比較することで2つのスタック領域の間の共存関係を調べることができる．2つのスタック領域は，図-2に示すように，スタック領域の両端に位置する水素結合を形成する塩基の番号（円内のアルファベット）が図中の不等式を満たした場合に共存できることになる．

3. スタック領域の抽出

2次構造を予測するために，まず，与えられた塩基列からスタック領域となり得る水素結合の連

AAAGCGCGCAAAAAAGCGCCGAAA
AAACGGCGCAAAAAAGGCGCCAAA
AAAGGCGCCAAAAAAGCGCCGAAA
AAACGGCGCAAAAAAGCGCGCAAA

(a) 塩基列



(b) 2つの安定構造

図-1 環状RNAの塩基列と2つの安定構造

† RNA Secondary Structure Prediction and Visualization using Highly Parallel Computers by Akihiro NAKAYA (Department of Information Science, Graduate School of Science, University of Tokyo.Presently with Systems Development Laboratory, Hitachi, Ltd.), Kenji YAMAMOTO (Bun'in-Hospital, Faculty of Medicine, University of Tokyo) and Akinori YONEZAWA (Department of Information Science, Graduate School of Science, University of Tokyo).

†† 東京大学大学院 理学系研究科 情報科学専攻．現在：（株）日立製作所 システム開発研究所
††† 東京大学医学部分院
†††† 東京大学大学院 理学系研究科 情報科学専攻

続部分（候補）を求める．2次構造全体のエネルギーは互いに共存可能なスタック領域のエネルギーの和として求めることができるため[5]，エネルギー和が最小化するようなスタック領域の集合を求めればよいことになる．候補はエネルギーによって整列，番号づけし，安定度が上位N（Nはパラメータ）のものを用いて探索を行う．表-1は図-1(a)の塩基列から取り出したスタック領域である．ここでは，Nは6とした．

4. 探索木の生成と枝刈

2次構造を求めるために深さiのノードがi番目の候補を入れるか否かを判断する探索木を生成する．探索木の左枝が候補を入れることに相当するとする．2次構造は探索木の葉の位置に得られる．この木の葉は2^N（Nは候補の数）だけあるため，次のような不必要な枝を効率よく刈る必要がある．(i)前述の共存条件によって生成することができない枝(ii)安定構造の得られる見込みのない枝である．見込みのない枝を刈るために枝の下に生じる安定構造のエネルギーの下限値Eを求める．また，E_tは時刻tまでに求められた最安定2次構造のエネルギーを記録したものであり，より安定な構造が得られるたびに更新されるものとする．各枝の下限値EをE_tと比較し，

$$E - E_t < \Delta K \qquad (1)$$

を満たしたときのみ枝を生成する（ΔKはパラメータ）．時刻tではE_tに比べてΔKkcalだけ不安定な解を生じる可能性がある枝は刈らないため，最終的に探索が終了した時点では，最安定解に比べてΔKkcalだけ不安定な解がもれなく求められて

表-1 スタック領域

候補	構成塩基	(−kcal)
0	^{40}GGCGC^{45}C・・・^{52}GGCGC^{57}C	19.9
1	^{28}CGGCG^{33}C・・・^{64}GCGCC^{69}G	19.4
2	^{16}GCGCC^{21}G・・・^{28}CGGCG^{33}C	19.4
3	^{16}GCGCC^{21}G・・・^{76}CGGCG^{81}C	19.4
4	^{64}GCGCC^{69}G・・・^{76}CGGCG^{81}C	19.4
5	^{4}GCGCG^{9}C・・・^{88}GCGCG^{93}C	17.6

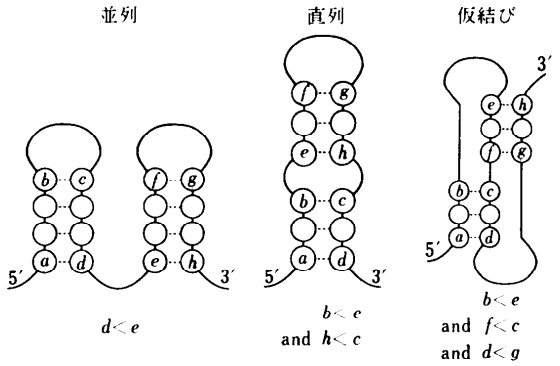


図-2 2つのスタック領域の共存関係

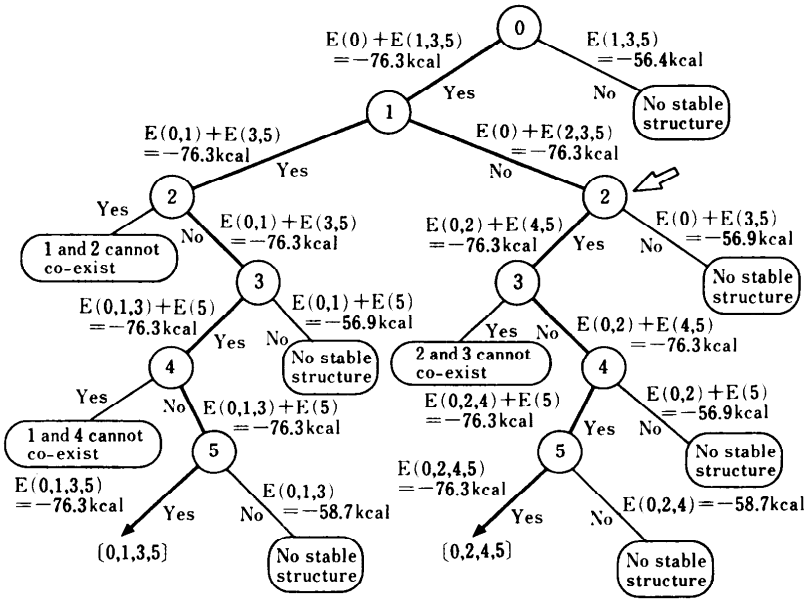


図-3 探索木の生成と枝刈

表-2 候補数と探索時間の関係

環境：候補数	95	150	203
CM-5	14秒	3分18秒	1時間01分
SS20	18秒	11分45秒	3時間48分
PC/AT	35秒	31分58秒	6時間50分

いることが保証される．$t=0$では，E_0は最安定解のエネルギー値を下回らない値を用いる．

枝の下に生じる安定構造のエネルギーの下限値Eを求めるために，incompatibility islets（islets）と呼ばれるデータ構造[6]を用いることができる．このisletsはすべての候補を共有部分のない部分集合に分割したものであり，各部分集合（islet）に属する任意の2つの候補は互いに共存できないという条件を満たすように構成される．

2次構造には各islet から高々1つの候補のみしか参加できないことを利用してエネルギーの評価と枝刈を行う．あるノードの下の枝に対する下限値は，根ノードからそのノードに至るまでに選ばれることが確定済みの候補と共存可能で，かつislet内で最も安定な候補を各isletから高々1つ選びそれらのエネルギー和として求めることができる．

たとえば，表-1のスタック領域のisletsとして次のようなリストIで表された候補の集合の分割を用いることができる（数字は候補の番号）．

$$I=(l_0,l_1,l_2,l_3)=((0),(1,2),(3,4),(5)) \quad (2)$$

仮に$E_0=-70.0$，$\Delta K=4.0$として探索を開始したとする．根ノードの左枝（候補0を入れることが確定）の下に生じる2次構造の下限値は次のように求められる．まず，l_0からは候補0の−19.9kcalが確定済み．l_1からは，候補1と候補2は共に候補0と共存可能でエネルギー値も等しく同条件であるが，たとえば，番号の小さい候補1を選んで−19.4kcal．同様にl_2からは候補3を選択して−19.4kcal．l_3からは候補5が選ばれて−17.6kcal．これらの4候補のエネルギーの和$E=-76.3$kcalが左枝の下限値である．これは式(1)を満たすのでこの枝は生成される．同様に，根ノードの右枝（候補0を入れないことが確定）の下限値は，候補1,3,5のエネルギー和（−56.4kcal）となるが，式(1)を満たさないので枝刈の対象となる．左右の枝の両方が生成される場合は下限値の小さい枝が先に生成される．以上のような判断を各ノードで繰り返した探索木の全体を図示したのが図-3である．図中，枝と共に示された値がその枝の下限値である．$E(x,y,...)$は候補$x,y,...$のエネルギー和を表し，"+"の左辺は確定した候補のエネルギー和，右辺は非確定候補のエネルギー和を表す．この例では2つの2次構造（図-1(b)）がそれぞれを構成する候補の番号のリストとして得られている．

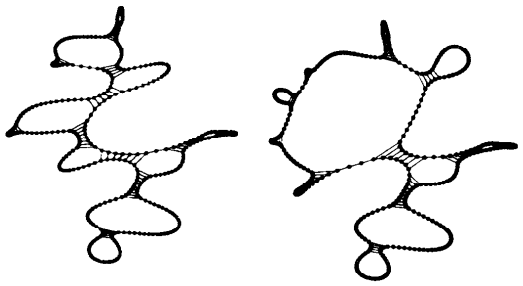


図-4 PSTVの2次構造

5. 探索の並列化

探索木中のノード，たとえば，図-3白い矢印のついたノードは，(i)根ノードからの経路を表すビット列10xxxx (ii)木中の深さ2(iii)下限値−76.3kcalという3つのデータによって表される（ビット列のiビット目は候補iが選ばれたときに1，選ばれないときは0が代入されているとし，未確定のときはxとする）．この3つのデータをまとめてタスクと呼ぶことにする．このタスクを他のPEに送付しリモートな枝を生成する．並列計算機上では，PE数と同等程度の数生成された部分木の根ノードに相当するタスクが各PEに割り当てられ（タスクの初期分配），並列に探索を開始する．PE上でローカルに行われる計算は逐次プログラムのそれと同じである．ところが，この初期分配によって各PEに割り振られる計算量は均一ではなく，枝刈によって仕事がなくなってしまうPEが生じてくる．このためPE間では動的なタスク分配が行われる．

各PE上で得られた2次構造のエネルギー値はローカルな変数E_l'によって保持される．あるPEで新しく2次構造が発見され，E_l'の値が更新される場合，その更新はすべてのPEにブロードキャストされる．この際，E_l'が更新されることによって

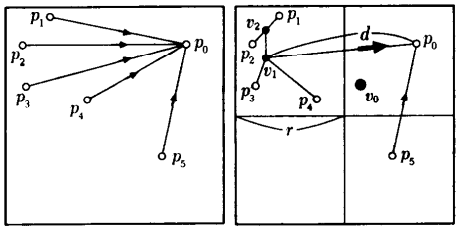


図-5 粒子群の重心に位置する仮想粒子の導入

枝刈が多くのPE上で行われることが期待できる．より安定な構造を早い段階で発見し，その結果をすべてのPEにブロードキャストすることによって，単に逐次探索を複数のPEで分担する以上の効果が得られる．

表-2はジャガイモやせいも病ウィロイド(PSTV:359塩基)に関する候補数と予測に要する計算時間を並列計算機と逐次計算機上（CM-5：SPARC(33MHz)×32，SS20：hyperSPARC(150MHz)，PC/AT：IntelDX4(100MHz)）で計測したものである．203候補を用いた探索では，合計31の2次構造を得た．その内2つは図-4のような形状をしている．プロセッサの能力差を考慮すれば，並列計算機上での探索は十分よい結果を得ているといえる．

6. 2次構造の視覚化

この章では，前章で求められた2次構造を塩基間の力にもとづいた2次元平面上の描画アルゴリズム[4),7)]について紹介する．まず，塩基間の引力と斥力を定義し，それにもとづいたナイーブなアルゴリズムについて述べる．さらに，BarnesとHutによる近似[8)]の導入とその並列化について述べる．

7. ナイーブなアルゴリズム

まず，粒子間の斥力と引力を定義する．物理的な力と異なり，ここでの「力」は単に粒子の収束位置からのズレを表す便宜的な指標としてのみ用いられており，物理的な法則には必ずしも合致しないものである．位置p_iをもつ粒子iに位置p_jをもつ粒子jが及ぼす斥力f^{ij}_{rep}および，位置p_jをもつ粒子jが位置p_iをもつ粒子iに及ぼす引力f^{ij}_{att}を次のように定義する．ここで，C_{rep}，C_{att}は定数．

$$f^{ij}_{\mathrm{rep}}=\frac{C_{\mathrm{rep}}}{|p_j-p_i|^2}\frac{-p_j+p_i}{|p_j-p_i|} \tag{3}$$

$$f^{ij}_{\mathrm{att}}=\begin{cases} C_{\mathrm{att}}(p_j-p_i) & \cdots \text{case 1} \\ 0 & \cdots \text{case 2} \end{cases} \tag{4}$$

ここで，case 1:p_jとp_iの間に水素結合があるか，隣同士の粒子であるとき．case 2:それ以外，とする．N粒子系の中でp_iが他のN-1粒子から受ける力fiは次のように求められる．

$$f_i=\sum_{k\neq i} f^{ik}_{\mathrm{rep}}+\sum_{k\neq i} f^{ik}_{\mathrm{att}} \tag{5}$$

各粒子は初期位置（たとえば円状）から始めて力の計算と位置の更新を繰り返していく．i番目の粒子の時刻tの位置を$p_{i,t}$とすると，微小時間Δt後の位置$p_{i,t+\Delta t}$は次のように求められる．すべての粒子は等しい質量mをもつものとする．

$$p_{i,t+\Delta t}=p_{i,t}+\Delta t f_i/m \tag{6}$$

8. 近似アルゴリズム

前述のナイーブなアルゴリズムでは，図-5に示すように1つの粒子に働く斥力の計算は$O(N)$を要するため，全体では$O(N^2)$の計算量となり，Nが大きくなったときには必ずしも実用的ではない．$O(N^2)$の計算量に対応するために，N体シミュレーションで用いられる近似アルゴリズムを斥力の計算に用いることができる．ここでは，BarnesとHutによる計算量$O(N\log N)$の近似アルゴリズムを用いる．このアルゴリズムの中心的アイデアは，ある粒子からみて「十分遠く」にある粒子群をそれらの重心に位置する仮想的な粒子（図-5）で置き換えることである（仮想粒子の質量は近似される粒子群の質量の和とする）．

ここで，v_0とv_1，v_2が仮想粒子であり，v_2はp_1とp_2の重心に，v_1はv_2とp_3，p_4の重心に位置する．v_0は全粒子を表す仮想粒子．このとき，図-5に示すように，p_0とv_1の距離をdとし，v_1として近似された粒子群が一辺rの正方形内（左上）に存在していたとする．このとき$r/d\leqq\theta$（θは定数）を満たしたならば，「十分遠く」であるとし，p_1,p_2,p_3,p_4はv_1として近似することができる．この近似によってp_0に及ぶ斥力はv_1とp_5から計算すればよいことになる．この近似の効果（θと計算時間の関係）は表-3の通りである．θの増加と共に近似の効果によって計算時間の短縮が実現さ

表-3 θと計算時間の関係.
計算時間は，逐次C++によって記述されたプログラムを用いて図-4に示された2次構造の視覚化に要したものである．「近似なし」は計算量$O(N^2)$のアルゴリズムに要した計算時間である．

θ	0	0.2	0.67	1.0	2.0	近似なし
秒	2475	648	210	144	92	481

れている．$\theta = 0, 0.2$では近似のためのコスト（木の生成等）が近似の効果を上回っている．

この近似を効率よく実現するために，粒子が配置する空間は分割され，4分木を用いて表される．空間に含まれる粒子が2個以上ある場合，4つの部分空間に分割される．すべての粒子が含まれる空間にこの操作を再帰的に施して分割された空間を得る．図-6は粒子の位置の更新の様子と空間の分割の様子を表している．

9. 描画の並列化

Barnes と Hutのアルゴリズムは空間の分割（木の生成）と粒子の位置の更新の2つの段階に分けられ，それぞれ次のように並列化できる．

木の生成：粒子が含まれる空間の分割は，空の木に粒子を1つずつ加えていくことで実現する．空間全体を表す根ノード（図-5のv_0）のみから始める．この根ノードは粒子を受け取ると，その座標に応じて，部分空間に加えていく．部分空間に含まれる粒子数が2以上になると，その空間はさらに分割され，仮想粒子が生成される（v_1,v_2）．このとき，粒子p_0, p_5,v_1に相当する木のノードが異なるPE上に分散して生成されていれば，v_0のこれら3つの部分木への粒子の追加は並列化することができる．同様に重心と質量の設定も異なるPEに存在する仮想粒子に対しては並列に行うことができる．

粒子の位置の更新：いったん，木が生成されてしまえば，すべての粒子の位置更新は互いに独立に（並列に）行うことができる．この段階で各粒子に働く力を計算するために分割された空間を表す木への参照が行われるが，前述のようにこの木はPE上に分散して生成されているため，木への参照は並列に行うことができる．

実装は並列オブジェクト指向言語 Schematic[9]を用いて行った．Schematic は Lisp の1方言である Scheme に並列オブジェクトの定義と，

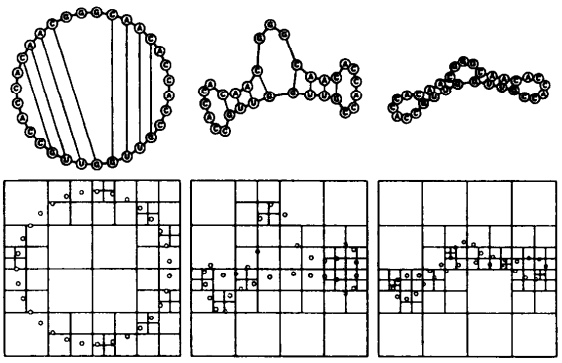


図-6 描画の過程と空間分割

Multilisp のfuture を変更／拡張したものを加えたものであり，分散メモリ型並列計算機およびワークステーションクラスタ上で動作している．

10. おわりに

我々は多数の RNA 2次構造を求めることや，N体問題を応用した2次構造の視覚化といった，計算量的に困難な問題を解くことを第1の目標にしてきた．次の目標として，計算結果からいかに有効な考察を引き出すかということを目的とする生物学者のために，クラスタ解析に基づく2次構造の自動的な分類[4]など補助的なツールの準備があげられる．また，一連のツール（予測，視覚化，分類）の何らかの形（フリーソフト化，商品化，ネットワーク経由でのサービスなど）での提供が期待されている．

一方で，対称相互結合型ニューラルネットワークを用いた手法[10]が提案されており，数千塩基のRNAの2次構造を我々の手法と同様にスタック領域を組み合わせることによって，非常に高速に求めることが可能となっている．また，その並列化も試みられている．しかし，準安定解もすべて求めるという我々のアルゴリズムとは性格を異にするものである．

参 考 文 献

1) Zucker, M. and Stiegler, P.: Optimal Computer Folding of Large RNA Sequence using Thermodynamics and Auxiliary Information, Nucleic Acids Res., Vol.9, pp.133-148 (1980).

2)中谷明弘, 田浦健次朗, 米澤明憲 : 並列オブジェクト指向言語ABCL/*f*によるRNA2次構造予測,情報処理

学会研究報告, 95-HPC-57, pp.25-30 (1995).
3) Nakaya, A., Yamamoto, K. and Yonezawa, A.:RNA Secondary Structure Prediction using Highly Parallel Computers, Comput. Applic.Biosci., Vol.11, No.6, pp.685-692 (1995).
4) Nakaya, A.: RNA Secondary Structure Prediction and Visualization using Highly Parallel Computers, Master's thesis, University of Tokyo (1996).
5) Salser, W.: Globin mRNA Sequences; Analysis of Base Pairing and Evolutionary Implications, Cold Spring Harbor Symp. Quant. Biol., Vol.42, pp.985-1002 (1977).
6) Dumas, J.-P. and Ninio, J.: Efficient Algorithms for Folding and Computing Nucleic Acid Sequences, Nucleic Acids Res., Vol.10, No.1, pp.197-206 (1982).
7) Nakaya, A., Taura, K., Yamamoto, K. and Yonezawa, A.: Visualization of RNA Secondary Structures using Highly Parallel Computers, Comput. Applic. Biosci. (1996).
8) Barnes, J. and Hut, P.: A Hierarchical O(NlogN) Force-calculation Algorithm, Nature, Vol.324, pp.446-449 (1986).
9) Taura, K. and Yonezawa, A.: Schematic: A Concurrent Object-oriented Extension to Scheme, Proceedings of Workshop on Object-Based Parallel and Distributed Computation, LNCS, Spring-Verlag (1996). (to appear).
10) 秋山 泰, 古谷立美:対称相互結合型ニューラルネットワークにおけるエネルギー極小化現象を利用した高速なRNA二次構造予測法, 1992年情報学シンポジウム論文集, 情報処理学会, pp.125-134(1992).

(平成8年6月24日受付)

中谷 明弘 (正会員)

1970年生. 1994年東京大学理学部情報科学科卒業. 1996年同大学院修士課程修了. 同年 (株) 日立製作所入社. 日本ソフトウエア科学会会員.

山本 健二

1954年生. 1976年東京大学理学部数学科卒業. 1985年同医学部医学科卒業. 同医学部分院勤務. 医学博士. 分子遺伝学に興味をもつ. 日本分子生物学会会員.

米澤 明憲 (正会員)

1947年生. 1977年 Ph.D. (MIT). 1989年より東京大学理学部情報科学科教授. 超並列ソフトウエアアーキテクチャ, ソフトウエア基礎論, などに興味をもつ. 共著書「算法表現論」,「モデルと表現」(岩波書店), 編著書「ABCL:An Object-Oriented Concurrent System」(MIT Press)などがある. 1992年より4年間ドイツ国立情報処理研究所 (GMD) 科学顧問. 現在 IEEEParallel & Distributed Technology編集委員, 日本ソフトウエア科学会理事長.