2007－MPS－66 (26)
2007／9／4

組合せ爆発を内包する化学反応系の平衡状態計算

小林 聡
電気通信大学 情報工学科

分子がさまざまな形で会合して複合分子を形成する化学反応系は，DNA 計算，DNA ナノテクノロジー，バイオインフォマティクスなどの分野で重要な研究対象である．DNA 計算・DNA ナノテクノロジーの分野では，DNA タイルや DNA 配列の会合反応，バイオインフォマティクスの分野では，RNA 配列の会合反応，たんぱく質の会合反応などはその例として挙げられる．このような反応系は，生成される分子の複合体の種類が組合せ爆発を起こすという問題を内包しており，現在まで，効率の良い解析手法は与えられていない．本論文は，このような組合せ爆発を内包する会合反応系の平衡状態を効率良く計算するための一般的な方法論を展開する．

Computing Equilibrium State of Chemical Reaction Systems Involving Combinatorial Explosion Problem

Satoshi Kobayashi
Dept. of Computer Science, Univ. of Electro-Communications
e-mail:satoshi@cs.uec.ac.jp

In the research areas, such as DNA computing, DNA nanotechnology, bioinformatics, etc., it is very important to study chemical reaction systems in which molecules are hybridizing in various ways to produce a number of complexes of molecules. For example, such reaction systems include a tile assembly system, DNA sequence hybridization system in DNA computing and DNA nanotechnology, and interacting RNA molecules in bioinformatics. These reaction systems involve a combinatorial explosion problem of the number of resultant complexes of molecules, which makes it difficult to obtain an efficient methodology for analyzing them. This paper proposes a general framework for efficiently computing equilibrium state of hybridization reaction systems which involve this combinatorial explosion problem.

1 はじめに

DNA 計算，DNA ナノテクノロジー，バイオインフォマティクスなどの分野では，分子がさまざまな形で会合することにより複雑な複合体を形成するような化学反応系を解析することの重要性が増している．DNA 計算・DNA ナノテクノロジーの分野では，DNA タイルや DNA 配列の会合反応，バイオインフォマティクスの分野では，RNA 配列の会合反応，たんぱく質の会合反応などはその例として挙げられる．本論文では，このように生成される複合分子の種類が組合せ爆発を引き起こすような会合反応系の平衡状態を，効率良く計算するための新しい枠組み・方法論を提案する．

この問題を解決する上で重要な鍵となる考え方に，複合分子がとる構造の自由エネルギー計算における「局所性」がある．例えば，RNA 二次構造では，ヘアピン，バルジ，内部ループといった局所的な構造がもつ自由エネルギーの総和を計算することにより，与えられた二次構造全体の自由エネルギーを計算できる．このような局所性は，動的計画法による二次構造予測アルゴリズムにおける重要な基礎となっている ([ZS81][Kob04])．このような構造の「局所性」を一般的にとらえる枠組みを提案するために，本論文では，「グラフによる複合分子の列挙」という新しい考え方を適用する．

二つめの重要な考え方は，「二段階最適化」という解き方である．平衡状態を求める問題は，物理学的には，反応系全体の自由エネルギーを最小化する問題に定式化することができる．本論文では，この最適化問題を直接解くのではなく，二段階に分けて解く．そのために，反応によって生成される複合分子の濃度分布の間に同値関係を導入する．すると，各同値類は濃度分布の集合となるが，第一段階で，各同値類の中で最適な濃度分布を解析的に求める．第二段階で，それらの局所的最適解の中から大局的な最適解を求める問題を解く．提案手法では，第二段階で凸計画法を適用する．

関連研究として，ごく最近，RNA 分子が会合する反応系の平衡状態を計算する手法が，Dirks，Bois らによって提案されている ([DBS07])．しかしながら，彼らの手法は希薄な溶液を仮定しており，平衡状態において構造分子が干渉し合わないという単純化を行っている．本論文では，そのような単純化を行わない一般の場合において，効率良く厳密解を得る方法を与える．また，さらに，本論文の手法は，RNA 二次構造の会合反応系だけでなく，さまざまな会合反応系に適用できる一般論を展開している点が大きく異なる．また，Adleman らもごく単純な一次元のタイルアセンブリに対する解析を行っているが ([ACG00])，本論文では，一般的な手法を与えており，より複雑で現実的な会合反応系に対して適用可能である点が

異なる．

2 平衡状態計算と自由エネルギー

数値の集合 V に対して，V_+ によって，非負の要素からなる V の部分集合を表す．また，V_{++} によって，正の要素からなる V の部分集合を表す．$\mathbf{R}$ を実数の集合，$\mathbf{Z}$ を整数の集合とする．

$\mathcal{M}$ を分子の有限集合，$\mathcal{A}$ を分子の複合体の有限集合とする．分子 $x \in \mathcal{M}$ と複合体 $X \in \mathcal{A}$ に対し，$\#(x, X)$ によって X に出現する x の個数を表す．$\mathcal{A}$ の要素からなる有限多重集合の対 $(\mathcal{X}_1, \mathcal{X}_2)$ で以下の条件を満たすものを反応規則という．

$$\sum_{X \in \mathcal{X}_1} \#(x, X) = \sum_{X \in \mathcal{X}_2} \#(x, X) \quad (\forall x \in \mathcal{M}) \quad (1)$$

この条件は保存則に対応する．反応規則 $(\mathcal{X}_1,\mathcal{X}_2)$ は通常 $\mathcal{X}_1 \rightleftharpoons \mathcal{X}_2$ と書かれる．$\mathcal{X}_1 = \{X_1, ..., X_{n_1}\}$ および $\mathcal{X}_1 = \{Y_1, ..., Y_{n_2}\}$ と書けるときは，

$$X_1 + \cdots + X_{n_1} \rightleftharpoons Y_1 + \cdots + Y_{n_2}.$$

と書かれる．

集合 $\mathcal{U}$ の濃度分布は $\mathcal{U}$ から $\mathbf{R}^+$ への関数と定義される．濃度分布を表すために，$[]$, $[]_1$, $[]_2$ などの記法を用いる．例えば，$\mathcal{A}$ の濃度分布 $[]$ と複合体 $X \in \mathcal{A}$ に対し，$[X]$ は X の濃度を表す．

反応の開始時点においては，分子の集合 $\mathcal{M}$ が初期濃度分布 $[]_0$ で提供されるものとする．すると，任意の時点において，反応系の $\mathcal{A}$ の濃度分布 $[]$ は以下の保存則を満たさねばならない．

$$\sum_{X \in \mathcal{A}} \#(x, X) \cdot [X] = [x]_0 \quad (\forall x \in \mathcal{M}) \quad (2)$$

平衡状態では，$\mathcal{A}$ の濃度分布 $[]$ は，各反応規則 $\mathcal{X}_1 \rightleftharpoons \mathcal{X}_2$ に対して，以下の平衡式を満たす．

$$e^{\sum_{X \in \mathcal{X}_1} E(X)} \times \prod_{X \in \mathcal{X}_1} [X] = e^{\sum_{X \in \mathcal{X}_2} E(X)} \times \prod_{X \in \mathcal{X}_2} [X] \quad (3)$$

ここで，$E(X)$ は複合体 $X \in \mathcal{A}$ の自由エネルギーを表す．

本論文では，上述したような会合反応系が与えられたとき，その平衡状態，つまり，すべての反応規則の平衡式 (3) とすべての分子 $x \in \mathcal{M}$ の保存則 (2) を満たすような $\mathcal{A}$ の濃度分布 $[]$ を求める問題を考察する．

会合反応系は，すべての分子 $x \in \mathcal{M}$ の保存則 (2) を満たすような $\mathcal{A}$ の濃度分布 $[]$ が存在するとき，**無矛盾である**という．$\mathcal{M} \subseteq \mathcal{A}$ を満たすような会合反応系は無矛盾であることが示せる．分子が最終反応物に含まれるというこの仮定は自然であり，これ以後は，$\mathcal{M} \subseteq \mathcal{A}$ を満たす反応系のみを考える．また，分子の初期濃度 $[]_0$ も，すべての分子 $x \in \mathcal{M}$ に対して，$[x]_0 > 0$ が成り立つものとする．

平衡状態を計算する問題は，反応系全体の自由エネルギーを考えることにより最適化問題に帰着できる．具体的には，$\mathcal{A}$ の濃度分布を $[]$ で表すと，反応系 P の全体の自由エネルギーは以下のように定義できる．

$$FE_1(P, []) = \sum_{X \in \mathcal{A}} E(X) \cdot [X] + \sum_{X \in \mathcal{A}} [X](\log[X] - 1). \quad (4)$$

平衡状態を計算するための制約付き最適化問題は以下で与えられる．

Free Energy Minimization Problem 1 (FEMP1)
minimize : $FE_1(P)$
subject to :

$$\sum_{X \in \mathcal{A}} \#(x, X) \cdot [X] = [x]_0, \quad (\forall x \in \mathcal{M})$$
$$[X] \geq 0. \quad (\forall X \in \mathcal{A})$$

ここで，$E(X)$ は単なる定数であり，変数は $[X]$ $(X \in \mathcal{A})$ である．この最適化問題と平衡状態との関係は，KKT 条件を用いて証明することにより，以下のように述べることができる．

Theorem 1 $\mathcal{A}$ の濃度分布 $[]$ が FEMP1 の最適解ならば，$[]$ は反応系 P の平衡状態である． □

$|\mathcal{A}| = n$ とおくと，$FE_1(P, [])$ は $\mathbf{R}^n_+$ 上で連続かつ凸であり，$\mathbf{R}^n_{++}$ 上で 2 階微分可能である．従って，FEMP1 は凸計画問題であり，凸計画法の理論を用いて平衡状態を求めることができることがわかる．しかしながら，本論文では，$\mathcal{A}$ の要素数が組合せ爆発を起こすような場合を想定している．つまり，FEMP1 のままでは，変数の個数が組合せ爆発を起こすので，凸計画法によって効率良く解を求めることはできない．本論文では，この問題を打開して，変数の個数を劇的に減らすための一般的な手法を与える．その際に，重要な鍵となる考えが，構造の「局所性」という考え方と，「二段階最適化」という解き方である．

3 構造の局所性

RNA 二次構造の自由エネルギーは，その局所構造 (ヘアピン，バルジループなど) の自由エネルギーの総和を求めることにより得られる．この性質のおかげで，二次構造予測問題を動的計画法を用いて効率良く求めることができる．このような局所性は，平衡状態を求める問題においても，おそらく何らかの意味で役立ってくれるであろうという考えから，本節では，局所性とは何かという問いかけを行う．そ

して，以下の「グラフによる複合分子の列挙」という考え方を提案する．

アサイクリックな有向グラフ $G=(V,Eg)$ を考える．ここで，V は節点の集合，Eg は有向辺の集合である．入る辺を持たない節点の集合を V_0，出る辺を持たない節点の集合を V_f で表す．V_0 の要素を**初期節点**，V_f の要素を**最終節点**とよぶ．各節点 $v\in V$ に対して，初期節点からの v への道と v から最終節点への道が存在するとき，G は**無駄がない**という．本論文では，以後，G は無駄がないものとする．

グラフ G の初期節点から最終節点への道の集合を $PT(G)$ で表す．グラフによる複合分子の列挙とは，$PT(G)$ から $\mathcal{A}$ への上への写像 ψ を考えることである．つまり，G の初期節点から最終節点への道を数え上げることにより複合体をすべて数え上げることができる．その場合，異なる道により同じ複合体を重複して数えることも条件付きで許される．その条件については，後で述べる．

このような数え上げを行うと，局所的な構造はどのように捉えられるであろうか．それは，G の辺として捉えられる．つまり，さまざまな複合体に共通して，ある局所構造が含まれるように，G のさまざまな道に共通して，ある辺が含まれているわけである．しかし，辺を局所構造として捉えるためには，局所構造を構成する分子の種類とその自由エネルギーを辺に対応させなければならない．従って，G は以下のような重みつきグラフにする必要がある．

G には 2 つの関数 (重み) が附随している．1 つは $\overline{E}$ という Eg から $\mathbf{R}$ への関数である．辺 $e\in Eg$ に対して，$\overline{E}(e)$ は辺 e が対応する局所構造の自由エネルギーを表す．2 つめの関数は，$\overline{\#}$ という $\mathcal{M}\times Eg$ から $\mathbf{Z}_+$ への関数である．分子 $x\in\mathcal{M}$ と辺 $e\in Eg$ に対して，$\overline{\#}(x,e)$ は e が対応する局所構造に含まれる x の個数を表す．ただし，同じ分子が連続する辺にまたがって現れるときは，重複しないように $\overline{\#}$ を定義する．例えば，DNA 配列 s がヘアピン，バルジなどの構造を形成しているときは，s の 5'-末端を含む局所構造においてのみ，$\overline{\#}(s,e)$ をカウントするようにすれば，重複して s を数えてしまうことはない．

以上により，「グラフによる複合分子の列挙」を形式的に述べる準備が整った．会合反応系 P，重みつきでアサイクリックで無駄がない有向グラフ G，$PT(G)$ から $\mathcal{A}$ への上への写像 ψ を考える．$\mathcal{S}=(P,G,\psi)$ は，以下の条件をすべての $\gamma\in PT(G)$ に対して満たすとき**列挙スキーム**であるという．

$$E(\psi(\gamma)) = \sum_{e\in Eg\,\mathrm{s.t.}\,e\in\gamma}\overline{E}(e)$$

$$\#(x,\psi(\gamma)) = \sum_{e\in Eg\,\mathrm{s.t.}\,e\in\gamma}\overline{\#}(x,e).$$

表記 $e\in\gamma$ は道 γ に辺 e が含まれていることを表す．

道 $\gamma\in PT(G)$ のランク n_γ を $n_\gamma=|\psi^{-1}(\psi(\gamma))|$ で定義する．また，列挙スキーム $\mathcal{S}$ のランク $n_\mathcal{S}$ を $\mathbf{max}\{n_\gamma\mid\gamma\in PT(G)\}$ で定義する．列挙スキーム $\mathcal{S}=(P,G,\psi)$ は以下の条件を満たすとき**対称的**であるという．

(1) 任意の $e\in Eg$ と $e\in\gamma_1,\gamma_2$ であるような任意の $\gamma_1,\gamma_2\in PT(G)$ に対して，$n_{\gamma_1}=n_{\gamma_2}$ が成り立つ．

(2) 任意の k $(1\le k\le n_\mathcal{S})$ に対して，$k-1$ 個の G の同型写像 $\phi_1,...,\phi_{k-1}$ が存在して，$n_\gamma=k$ となる任意の道 $\gamma\in PT(G)$ に対して，以下が成り立つ．

$$\{\gamma,\phi_1(\gamma),...,\phi_{k-1}(\gamma)\}=\psi^{-1}(\psi(\gamma))$$

上記で，ϕ_i が同型写像であるというとき，対応する辺の $\overline{Eg}$ や $\overline{\#}$ の値も等しいということが要求される．

次節において，対称的な列挙スキームをもつような会合反応系に対して，平衡状態を効率良く計算する方法を与える．

1 つ 1 つの例がそれなりの考察を必要とするので，紙面の都合上，例を挙げることは省略するが，タイルの 1 次元アセンブリや，RNA や DNA 配列の線形な二次構造形成など，さまざまな会合反応系に対して，対称的な列挙スキームを構成することができる．また，このグラフによる列挙は，ハイパーグラフによる列挙に自然な形で拡張することができる．これにより，木構造の複合体を形成するような会合反応系や，シュードノットを含まない RNA や DNA の二次構造形成を行うような会合反応系に対して，対称的な列挙スキームを構成することもできるようになる．

4 二段階最適化法

本節では，以下の条件を満たす会合反応系 P に対する平衡状態計算方法を与える．

(A) P は対称的列挙スキーム $\mathcal{S}=(P,G,\psi)$ を持つ．

整数 $k=2,...,n_\mathcal{S}$ を考える．Θ_k によって，ランク k の道の集合に関して $\mathcal{S}$ が対称的であることを保証するための G の同型写像の集合を表す．また，$\Theta=\cup_{k=1}^{n_\mathcal{S}}\Theta_k$ と定義する．

便宜上，しばしば，$\psi(\gamma)$ の代わりに X_γ と書く．また，$X\in\mathcal{A}$ に対して，$PT(X)=\psi^{-1}(X)$ と定義する．辺 $e\in Eg$ に対して，e を含む任意の $\gamma\in PT(G)$ を選んで，$n_e=n_\gamma$ と定義する．$\mathcal{S}$ が対称的であるので，この定義は well-defined である．

$\mathcal{A}$ の濃度分布 $[\,]$ を考える．辺 $e\in Eg$ に対して，以下を定義する．

$$\overline{[e]} \stackrel{def}{\equiv} \sum_{\gamma\in PT(G)\,\mathrm{s.t.}\,e\in\gamma}\frac{[X_\gamma]}{n_\gamma}$$

直感的に述べると，$\overline{[e]}$ は e が対応する局所構造の濃度を表す．

濃度分布 $[]_1$ と $[]_2$ は，任意の $e \in Eg$ に対して $\overline{[e]}_1 = \overline{[e]}_2$ となるとき等価であるといい，$[]_1 \overset{lc}{\equiv} []_2$ と書く．

これにより，濃度分布の間に同値関係 $\overset{lc}{\equiv}$ が導入された．次に，各同値類における FEMP1 の最適解を求めることを考える．

同値関係 $\overset{lc}{\equiv}$ の同値類を考える．つまり，任意の辺 $e \in Eg$ に対して，$\overline{[e]} = w_e$ を満たすような濃度分布 $[]$ の集合を考える．ここで，w_e $(e \in Eg)$ は正の実定数である．また，w_e は以下の条件を満たすものと仮定する．

(C1) $$\forall v \in V - V_0 - V_f, \quad \sum_{e \in v_{in}} w_e = \sum_{e \in v_{out}} w_e.$$

(C2) $$\forall \theta \in \Theta_k \, \forall e \in Eg \text{ s.t. } n_e = k, \quad w_e = w_{\theta(e)}$$

ここで，各節点 v に対して，v_{out} (v_{in}) は v から出る (に入る) 辺の集合を表す．

便宜上，$w_v = \sum_{e \in v_{out}} w_e$ とする．$PT(G)$ の濃度分布 $[]_{+,G}$ を以下によって定義する．

$$[\gamma]_{+,G} = \frac{\prod_{e \in \gamma} w_e}{\prod_{v \in \gamma,\, v \notin V_0,\, v \notin V_f} w_v}, \tag{5}$$

さらに，$\mathcal{A}$ の濃度分布 $[]_*$ を $[]_{+,G}$ を用いて以下のように定義する．

$$[X]_* = \sum_{\gamma \in PT(X)} [\gamma]_{+,G} \tag{6}$$

ここで，定数 w_e によって指定される同値類において FEMP1 を解く問題を考える．これを FEMP2 とする．この FEMP2 の最適解は，KKT 条件等を用いることにより，式 (5) および式 (6) で定義される $[]_*$ で与えられることが示せる．そこで，この最適解を $FE_1(P)$ に代入して，以下の最適化問題を導入する．ただし，以下において，節点集合 W $(\subseteq V - V_f)$ に対して，$W_{out} = \cup_{v \in W} v_{out}$ と定義する．

Free Energy Minimization Problem 3 (FEMP3)
minimize :

$$FE_3(P, (w_e \mid e \in Eg)) \overset{def}{\equiv} \sum_{e \in Eg} \overline{E}(e) \cdot w_e + \sum_{e \in Eg} w_e(\log w_e - 1) - \sum_{v \in V - V_0 - V_f} w_v(\log w_v - 1) + \sum_{e \in (V_0)_{out}} w_e \cdot \log n_e$$

subject to :

$$\begin{aligned} \sum_{e \in Eg} \overline{\#}(x, e) \cdot w_e &= [x]_0, && (\forall x \in \mathcal{M}) \\ \sum_{e \in v_{in}} w_e &= \sum_{e \in v_{out}} w_e, && (\forall v \in V - V_0 - V_f\}) \\ w_e &= w_{\theta(e)}, && (\forall \theta \in \Theta_k \, \forall e \in Eg \text{ s.t. } n_e = k) \\ w_e &\geq 0. && (\forall e \in Eg) \end{aligned}$$

FEMP3 における変数は w_e $(e \in Eg)$ であることに注意されたい．従って，FEMP3 では変数の個数が $|\mathcal{A}|$ から $|Eg|$ に劇的に削減できている．そして，FEMP3 を解くことによって FEMP1 を解けることが以下のように示せる．

Theorem 2 **(A)** を仮定する．また，FEMP3 の最適解を $(w_e \mid e \in Eg)$ とする．この $(w_e \mid e \in Eg)$ に基づいて式 (5) と式 (6) によって定義される濃度分布 $[]_*$ は FEMP1 の最適解を与える． □

また，幸いなことに，FEMP3 の目的関数は $\mathbf{R}^m_{++}$ 上で凸関数であることが示せる．ここで，$m = |Eg|$ である．

よって，条件 **(A)** を満たす反応系は，FEMP3 を凸計画法 ([NN93]) を用いて解くことにより，効率良く解くことができる．

References

[ACG00] L. Adleman, Q. Cheng, A. Goel, M. Huang, H. Wasserman, Linear Self-Assemblies: Equilibria, Entropy, and Convergence Rates, unpublished manuscript, 2000.

[DBS07] R.M. Dirks, J.S. Bois, J.M. Schaeffer, E. Winfree, N.A. Pierce, Thermodynamic Analysis of Interacting Nucleic Acid Strands, *SIAM Review*, **49**, pp.65-88, 2007.

[Kob04] S. Kobayashi, Testing Structure Freeness of Regular Sets of Biomolecular Sequences, in *Preliminary Proceedings of 10th International Meeting on DNA Based Computers*, pp.395-404, 2004.

[NN93] Y. Nesterov and A. Nemirovskii, Interior-Point Polynomial Algorithms in Convex Programming, SIAM Studies in Applied and Numerical Mathematics, SIAM, Philadelphia, 1993.

[ZS81] M. Zuker, P. Steigler, Optimal Computer Folding of Large RNA Sequences using Thermodynamics and Auxiliary Information, *Nucleic Acids Research*, **9**, pp.133-148, 1981.