A New Approach to Computing Equilibrium State of Combinatorial Hybridization Reaction Systems [*]

Satoshi Kobayashi
Dept. of Computer Science, Univ. of Electro-Communications
1-5-1, Chofugaoka, Chofu, Tokyo 182-8585, JAPAN
E-mail: satoshi@cs.uec.ac.jp

ABSTRACT

This paper provides a new approach to the efficient analysis of equilibrium state of a combinatorial Hybridization Reaction System (HRS, for short) in which exponentially many assemblies of molecules are generated from a set of molecules. The proposed framewok provides a method to overcome the combinatorial explosion problem of resultant assemblies. The key idea exists in the *locality* of HRSs. The goal of this paper is to present a *general* theory for the efficient computation of equilibrium states.

Keywords

molecular computing, equilibrium, convex programming, hypergraphs, symmetry, tile assembly, DNA, RNA, secondary structure

1. INTRODUCTION

Since the pioneering work by Adleman([1]), the paradigm of *DNA computing* (in a broad sense, *molecular computing*) has emerged and attracted much attention from computer scientists, molecular biologists, DNA nanotechnologists, etc. The principle of DNA computing paradigm essentially relies on DNA hybridization process, but it is in essence error-prone. Therefore, it is substantially important to design a set of DNA sequences or tiles with which we can obtain a maximum concentration of a target molecular architecture. In order to evaluate a given set of DNA sequences or tiles in this respect, we need to devise a methodology for efficiently computing the concentration of the target assembly at the equilibrium state of hybridization reaction systems (HRSs, for short).

This paper gives a new approach to the efficient analysis of equilibrium state of HRSs by overcoming the combinatorial explosion problem of resultant assemblies. The key idea exists in the *locality* of HRSs. The goal of this paper is to present a *general* theory of the efficient computation of equilibrium states. Thus, it can be applied to various kinds of HRSs other than those of DNA and RNA molecules. As far as the author's knowledge, this is the first attempt to formulate such a general theory for computing equilibrium state of combinatorially complex hybridization reaction systems.

By locality, we intuitively mean the physical property that the free energy of an assembly X of molecules can be computed as the sum of free energies of all local substructures of X. For instance, the free energy of a single RNA and DNA molecule at the secondary structure level can be calculated as the sum of free energies of all local substructures such as hairpin loops, bulge loops, internal loops, multiple loops, etc. But, how we can formulate the concept of this kind of *locality* of HRSs in a *general* setting?

We will give a theoretical formulation of locality of HRSs using graph theory. We first require such an HRS of high locality to have the following properties: (1) there exists a weighted directed hypergraph G with initial and final vertices such that the set of hyperpaths from initial vertices to final vertices is in many-to-one correspondence with the set of assemblies of molecules, (2) the weight of a hyperpath is equivalent to the free energy of its corresponding assembly, (3) the hypergraph G has some symmetric structures which capture the symmetric property of the space of assemblies. In this formulation, a hyperarc of a hyperpath can be regarded as a substructure of its corresponding assembly. With this key concept of locality of HRSs, we will establish a general theory for computing equilibrium state of HRSs[1].

After providing the definition of HRSs, we will give a fundamental and preliminary discussion on the close relationship between the computation of equilibrium state and minimum free energy of HRSs in section 2. So, the problem of

[*] Earlier version of the results were reported in a technical report [6]. This work was supported in part by Grant-in-Aid for Scientific Research on Priority Area No.14085205, Ministry of Education, Culture, Sports, Science and Technology, Japan, and is supported by Grant-in-Aid for Scientific Research (C) NO.17500192, Japan Society for the Promotion of Science.

Bionetics '07, December 10-13, 2007, Budapest, Hungary
Copyright 2007 ICST 978-963-9799-11-0

[1] The theory proposed in this paper allows us to deal with globally defined entropic factors, such as those related to rotational symmetry of the assemblies, by decomposing graphs into subcomponents so that every component enumerates only assemblies with the same global entropic constant with respect to the concentration.

computing equilibrium state is converted to that of computing minimum free energy. Section 3 gives the definition of local HRSs. With this key concept of locality, in section 4, we will give a general theory of how to transform a free energy minimization problem with huge number of assemblies of molecules to that with a small number of variables.

2. EQUILIBRIUM STATE OF HYBRIDIZATION REACTION SYSTEM

For a set N of numbers, by N_+ and N_{++}, we denote the subsets of N consisting of only nonnegative and positive numbers, respectively. By $\mathbf{R}$ and $\mathbf{Z}$, we denote the set of real numbers and integers, respectively.

Let $\mathcal{M}$ be a finite set of *molecules* and $\mathcal{A}$ be a set of *assemblies of molecules* consisting of molecules in $\mathcal{M}$. For $x \in \mathcal{M}$ and $X \in \mathcal{A}$, by $\#(x, X)$, we denote the number of molecules x contained in an assembly X. A *reaction rule over* $\mathcal{A}$ is given by a pair of $\mathcal{X}_1$ and $\mathcal{X}_2$ of finite multisets consisting of elements of $\mathcal{A}$ such that the following equation holds:

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

Note that the sum over a multiset counts elements duplicatedly for their multiple occurrences. This equality constraint (1) corresponds to the law of conservation of each molecule. A reaction rule ($\mathcal{X}_1$,$\mathcal{X}_2$) is usually denoted by $\mathcal{X}_1 \rightleftharpoons \mathcal{X}_2$. In case of $\mathcal{X}_1 = \{X_1, ..., X_{n_1}\}$ and $\mathcal{X}_2 = \{Y_1, ..., Y_{n_2}\}$, where multiple occurrences of assemblies are allowed, we often write:

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

A *distribution* of a set $\mathcal{U}$ is a function from $\mathcal{U}$ to $\mathbf{R}_+$. Usually, we use notations, $[\,]$, $[\,]_1$, $[\,]_2$, ..., etc., for representing distributions. For example, for a distribution $[\,]$ of $\mathcal{A}$ and an assembly $X \in \mathcal{A}$, $[X]$ represents a concentration of the assembly X. A *distribution* of $\mathcal{M}$ is especially called an *initial distribution.* If we have a set $\mathcal{M}$ of molecules with its initial distribution $[\,]_0$, then any distribution $[\,]$ of $\mathcal{A}$ should satisfy the following equation:

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

This equality constraint corresponds to the law of conservation of each molecule.

For instance, let us consider two molecules α and β, and an assembly $\alpha\beta$ consisting of molecules α and β. Note that each of α and β is itself an assembly consisting of only one molecule. Thus, we can consider a reaction rule $\alpha + \beta \rightleftharpoons \alpha\beta$. Equilibrium state of this reaction rule is determined by free energies $E(\alpha)$, $E(\beta)$ and $E(\alpha\beta)$ of assemblies α, β, and $\alpha\beta$, respectively. More precisely, the distribution $[\,]$ at the equilibrium state should satisfy[2]:

$$\frac{[\alpha\beta]}{[\alpha][\beta]} \quad = \quad e^{-(E(\alpha\beta)-(E(\alpha)+E(\beta)))}$$

We will give the definition of this kind of *equilibrium equation* in a general setting. For a reaction rule $\mathcal{X}_1 \rightleftharpoons \mathcal{X}_2$, its equilibrium equation is given by:

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

In summary, a hybridization reaction system (HRS, for short) is defined by $P = (\mathcal{M}, \mathcal{A}, \#, \mathcal{R}, E, [\,]_0)$, where $\mathcal{M}$ is a nonempty set of molecules, $\mathcal{A}$ is a nonempty set of assemblies consisting of molecules in $\mathcal{M}$, $\#$ is a function from $\mathcal{M} \times \mathcal{A}$ to $\mathbf{N}_+$ such that $\#(x, X)$ indicates the number of molecules x contained in an assembly X, $\mathcal{R}$ is a set of reaction rules satisfying the equations (1), E is a free energy function from $\mathcal{A}$ to $\mathbf{R}$, and $[\,]_0$ is an initial distribution of $\mathcal{M}$. In the rest of this paper, we assume that $\mathcal{M}$, $\mathcal{A}$ and $\mathcal{R}$ are finite, $\mathcal{M} \subseteq \mathcal{A}$, and $[x]_0 > 0$ holds for every $x \in \mathcal{M}$.

The problem of interest is to find an equilibrium state of P, i.e., a distribution $[\,]$ of $\mathcal{A}$ satisfying equilibrium equations (3) of all $r \in \mathcal{R}$ and conservation laws (2) of all molecules $x \in \mathcal{M}$. Such a distribution $[\,]$ is called an *equilibrium state* of P.

For instance, let us define an HRS for the above example reaction $\alpha+\beta \rightleftharpoons \alpha\beta$. Consider an HRS $P = (\mathcal{M}, \mathcal{A}, \#, \mathcal{R}, E, [\,]_0)$, where $\mathcal{M} = \{\alpha, \beta\}$, $\mathcal{A} = \{\alpha, \beta, \alpha\beta\}$, $\mathcal{R} = \{\alpha + \beta \rightleftharpoons \alpha\beta\}$, and a function $\#$ is defined by: $\#(\alpha, \alpha) = 1$, $\#(\alpha, \beta) = 0$, $\#(\alpha, \alpha\beta) = 1$, $\#(\beta, \alpha) = 0$, $\#(\beta, \beta) = 1$, $\#(\beta, \alpha\beta) = 1$.

Then, the problem is to find a distribution $[\,]$ of $\mathcal{A}$ satisfying:

$$e^{E(\alpha\beta)} \times [\alpha\beta] = e^{E(\alpha)+E(\beta)} \times [\alpha][\beta],$$
$$[\alpha] + [\alpha\beta] = [\alpha]_0, \quad [\beta] + [\alpha\beta] = [\beta]_0.$$

Let $P = (\mathcal{M}, \mathcal{A}, \#, \mathcal{R}, E, [\,]_0)$ be an HRS. The *free energy* $FE_1(P, [\,])$ of P under distribution $[\,]$ of $\mathcal{A}$ is defined by:

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

where we define $0 \log 0 = 0$. Note that for any $X \in \mathcal{A}$, $E(X)$ is regarded as a constant. Free energy $FE_1(P, [\,])$ can be regarded as a function with respect to the *variables* $[X]$'s ($X \in \mathcal{A}$). We often simply write $FE_1(P)$ instead of $FE_1(P, [\,])$ if the context allows.

Consider the following minimization problem:

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

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

The following theorem is a well-known result (but, the author does not know who is the first one who found it).

THEOREM 1. *A distribution* $[\,]$ *of* $\mathcal{A}$ *is an equilibrium state of* P *if* $[\,]$ *is a minimizer of FEMP1.* □

Therefore, the problem of computing equilibrium state can be reduced to FEMP1. But, in this paper, we are interested in the case that the cardinality of $\mathcal{A}$ is tremendously larger than that of $\mathcal{M}$. We will give an approach to overcome such a difficulty.

3. LOCALITY OF HRS

[2] In this paper, the free energy $E(X)$ of X is a dimensionless quantity, i.e., $E(X)$ is the free energy per mol of X divided by the physical quantity RT, where R is the gas constant and T is the absolute temperature of the reaction system.

3.1 Hypergraphs

Basic notions and definitions related to directed hypergraphs will be introduced in this subsection mainly based on [4], but some notions are slightly different from its originals.

A *directed hypergraph* G is a pair (V, Eg), where V is a finite set of *vertices*, and Eg is a finite set of *hyperarcs* associated with two functions $t : Eg \rightarrow V$ and $H : Eg \rightarrow 2^V$. A directed hypergraph is simply referred as a *hypergraph* in this paper. A hyperarc e is interpreted as an arrow from a *tail* $t(e)$ to the set $H(e)$ of *heads* [3]. In this definition, we allow multi-hyperarcs, i.e. there can be more than one distinct hyperarcs with the same heads and a tail. For a vertex v, a hyperarc e such that $v = t(e)$ $(v \in H(e))$ is called an *outgoing* (*entering*) hyperarc of v. For a vertex v, by v_{out} (v_{in}), we denote the set of outgoing (entering) hyperarcs of v. For a set W of vertices, we define $W_{out} = \cup_{v\in W}\ v_{out}$ and $W_{in} = \cup_{v\in W}\ v_{in}$. By V_0 and V_f, we denote the set of vertices $v \in V$ such that $v_{in} = \emptyset$ and $v_{out} = \emptyset$, respectively. Elements of V_0 and V_f are called *initial vertices* and *final vertices*, respectively.

A *path from* s *to* u in G is a sequence $s = v_1, e_1, v_2, e_2, ..., e_q, v_{q+1} = u$ of vertices v_i $(i = 1, ..., q + 1)$ and hyperarcs e_i $(i = 1, ..., q)$ such that $v_i = t(e_i)$ and $v_{i+1} \in H(e_i)$ for $i = 1, ..., q$. If $s \in H(e_q)$ holds, the path is called a *cycle*. We say that G is *acyclic* if it contains no cycles.

A hypergraph $G' = (V', Eg')$ is called a *sub-hypergraph* of $G = (V, Eg)$ if $V' \subseteq V$ and $Eg' \subseteq Eg$ hold. Let x be an element of $V \cup Eg$. For a sub-hypergraph $G' = (V', Eg')$, we write $x \in G'$ if $x \in V' \cup Eg'$ holds. For a subset W of $V \cup Eg$, we write $W \subseteq G'$ if $x \in G'$ holds for every $x \in W$.

Let $r \in V$ and $S \subseteq V$. A *hyperpath of* G *from the root* r *to the sink set* S is a minimal acyclic sub-hypergraph γ of G such that r and S are contained in γ and every vertex of γ, except for those in S has exactly one outgoing hyperarc. A hyperpath is said to be *empty* if it contains only one vertex and no arcs (i.e., $S = \{r\}$). A hyperpath is said to be *elementary* if every vertex, except for r, has exactly one entering hyperarc. For a hypergraph $G = (V, Eg)$, by $PT(G)$, we denote the set of all hyperpaths from some root $r \in V_0$ to some sink set S with $S \subseteq V_f$. The hypergraph G is said to be *elementary* if every hyperpath in $PT(G)$ is elementary. We say that G is *reduced* if every hyperpath in $PT(G)$ is not an empty hyperpath.

Let ϕ be an injective and surjective mapping from $V \cup Eg$ to $V \cup Eg$ such that $\phi(V) = V$ and $\phi(Eg) = Eg$ hold [4]. If ϕ satisfies $\phi(t(e)) = t(\phi(e))$ and $\phi(H(e)) = H(\phi(e))$ for any $e \in Eg$, ϕ is called a *proper isomorphism* of G. On the other hand, if ϕ satisfies $\{\phi(t(e))\} = H(\phi(e))$ and $\phi(H(e)) = \{t(\phi(e))\}$ for any $e \in Eg$, ϕ is called an *anti-isomorphism* of G. Note that only ordinary graph, i.e., a hypergraph such that $|H(e)| = 1$ for every $e \in Eg$, can have an anti-isomorphism. We say that ϕ is an *isomorphism* of G if it is either a proper or an anti- isomorphism of G. Note that $V_0 = \phi(V_0)$ and $V_f = \phi(V_f)$ hold for a proper isomorphism ϕ of G. In case that ϕ is anti-isomorphism, we have $V_0 = \phi(V_f)$ and $V_f = \phi(V_0)$. For a subgraph $G' = (V', Eg')$ of G and an isomorphism ϕ of G, by $\phi(G')$ we denote the subgraph $(\phi(V'), \phi(Eg'))$.

3.2 Locality of HRS

Let $P = (\mathcal{M}, \mathcal{A}, \#, \mathcal{R}, E, [\,]_0)$ be an HRS and consider a reduced acyclic elementary hypergraph $G = (V, Eg)$ associated with two functions $\overline{E} : Eg \rightarrow \mathbf{R}$ and $\overline{\#} : \mathcal{M} \times Eg \rightarrow \mathbf{Z}_+$. An isomorphism ϕ of G is said to be *symmetric* if the following conditions are satisfied: (1) $\overline{E}(e) = \overline{E}(\phi(e))$ for all $e \in Eg$, (2) $\overline{\#}(x, e) = \overline{\#}(x, \phi(e))$, for all $x \in \mathcal{M}$ and $e \in Eg$.

Let ψ be a surjective function from $PT(G)$ to $\mathcal{A}$. Then, we say that a triple $\mathcal{S} = (P, G, \psi)$ is an *enumeration scheme* if the functions $\overline{E}$ and $\overline{\#}$ satisfy the following conditions for each $\gamma \in PT(G)$ and $x \in \mathcal{M}$:

$$\begin{aligned} E(\psi(\gamma)) &= \sum_{e\in Eg\,\mathrm{s.t.}\; e\in\gamma} \overline{E}(e), \\ \#(x, \psi(\gamma)) &= \sum_{e\in Eg\,\mathrm{s.t.}\; e\in\gamma} \overline{\#}(x, e). \end{aligned}$$

Let $\mathcal{S} = (P, G, \psi)$ be an enumeration scheme. The *rank* n_γ of $\gamma \in PT(G)$ is defined as $n_\gamma = |\psi^{-1}(\psi(\gamma))|$. The *rank set* $n_{\mathcal{S}}$ of $\mathcal{S}$ is defined as $n_{\mathcal{S}} = \{n_\gamma \mid \gamma \in PT(G)\}$.

An enumeration scheme $\mathcal{S} = (P, G, \psi)$ is said to be *symmetric* if:

(1) for any $e \in Eg$ and any $\gamma_1, \gamma_2 \in PT(G)$ with $e \in \gamma_1, \gamma_2$, $n_{\gamma_1} = n_{\gamma_2}$ holds, and

(2) for any $k \in n_{\mathcal{S}}$, there exist symmetric isomorphisms $\phi_1, ..., \phi_{k-1}$ of G such that for any $\gamma \in PT(G)$ with $n_\gamma = k$,

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

holds.

We can construct symmetric enumeration schemes for various HRSs dealing with one dimensional tile assembly, assembly of tree-like structures, RNA/DNA hybridization reactions, etc. In particular, the application to DNA/RNA secondary structures consisting of multiple sequences is important in DNA computing. For example, consider a class of linear secondary structures of multiple sequences ([5]) consisting of hairpin, internal, bulge loops and stacked base pairs. Each linear secondary structure can be determined by a sequence of base pairs $((\alpha_i, k_i), (\beta_i, l_i))$ $i = 1, ..., n$, where α_i and β_i are sequences, and k_i and l_i are the base positions of α_i and β_i, respectively. A pair $((\alpha_i, k_i), (\beta_i, l_i))$ indicates a base pair between k_ith base of α_i and l_ith base of β_i. If we regard a pair $((\alpha_i, k_i), (\beta_i, l_i))$ as a symbol, then the enumeration of linear secondary structure is accomplished by the enumeration of strings over such a base pair alphabet, although we need to consider symmetric property of such enumeration scheme. This idea can also be extended to pseudoknot free secondary structures consisting of multiple sequences. But, in this case, we need impose some restriction on the number of branches of multiloop substructures. In this way, we can apply the framework of symmetric enumeration sckem to various hybridization reaction systems, but because of space constraint, the details will be presented at the workshop.

4. REDUCING NUMBER OF VARIABLES

The difficulty for solving FEMP1 is that the cardinality of $\mathcal{A}$ is very large in real applications. In this paper, we will

[3] In its original definition, a hyperarc has a head and a set of tails.

[4] We need a mapping from $V \cup Eg$ to $V \cup Eg$ to define the concept *isomorphism*, because we allow multi-hyperarcs in G.

give a novel method to reduce the number of variables of FEMP1 in case that the following assumptions hold:

(A1) An HRS P to be investigated has an enumeration scheme $\mathcal{S} = (P, G, \psi)$.

(A2) The enumeration scheme $\mathcal{S} = (P, G, \psi)$ in **(A1)** is symmetric.

Let $P = (\mathcal{M}, \mathcal{A}, \#, \mathcal{R}, E, [\,]_0)$ be an HRS, $G = (V, Eg)$ be a reduced acyclic elementary hypergraph associated with two functions $\overline{E}$ and $\overline{\#}$, and ψ be a mapping from $PT(G)$ to $\mathcal{A}$ such that $\mathcal{S} = (P, G, \psi)$ is a symmetric enumeration scheme. For $k \in n_{\mathcal{S}}$ with $k \geq 2$, by Θ_k, we denote the set of all symmetric isomorphisms of G which are used to guarantee that $\mathcal{S}$ is symmetric with respect to the set of hyperpaths of rank k. By $\tilde{\Theta}_k$, we denote the minimal set of symmetric isomorphisms of G containing Θ_k and closed under composition and inverse. Define $\Theta = \cup_{k=2}^{n_{\mathcal{S}}} \Theta_k$ and $\tilde{\Theta} = \cup_{k=2}^{n_{\mathcal{S}}} \tilde{\Theta}_k$.

For convenience, we often write X_γ instead of $\psi(\gamma)$. For $X \in \mathcal{A}$, we define $PT(X) = \psi^{-1}(X)$ and $n_X = |PT(X)|$. For $e \in Eg$, we define $n_e = n_\gamma$ for some γ such that $e \in \gamma$. For $v \in V - V_0 - V_f$, we define $n_v = n_\gamma$ for some γ such that $v \in \gamma$. These definitions are well defined since the condition (1) holds in the definition of symmetric enumeration scheme.

Let $[\,]$ be a distribution of $\mathcal{A}$. For a vertex v and a hyperarc e of G, we define:

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

Intuitively speaking, $\overline{[e]}$ represents the concentration of the local structure corresponding to e.

PROPOSITION 1. *Assume* **(A1)**. *Let* $[\,]$ *be a distribution of* $\mathcal{A}$. *Then, we have* $\overline{[v]} = \sum_{e \in v_{out}} \overline{[e]}$ *for* $v \in V - V_f$, *and* $\overline{[v]} = \sum_{e \in v_{in}} \overline{[e]}$ *for* $v \in V - V_0$. □

PROPOSITION 2. *Assume* **(A1)** *and* **(A2)**. *Let* $[\,]$ *be a distribution of* $\mathcal{A}$. *We have* $\overline{[\theta(e)]} = \overline{[e]}$ *for every* $e \in Eg$ *and* $\theta \in \tilde{\Theta}_k$ *with* $n_e = k$. □

For distributions $[\,]_1$ and $[\,]_2$, we say that they are *locally equivalent*, written $[\,]_1 \stackrel{lc}{\equiv} [\,]_2$, if for any hyperarc e of G, $\overline{[e]}_1 = \overline{[e]}_2$ holds.

PROPOSITION 3. *Assume* **(A1)**. *The relation* $\stackrel{lc}{\equiv}$ *over the distributions of* $\mathcal{A}$ *is an equivalence relation.* □

PROPOSITION 4. *Assume* **(A1)**. *It holds that* $\sum_{X \in \mathcal{A}} E(X) \cdot [X] = \sum_{e \in Eg} \overline{E}(e) \cdot \overline{[e]}$. □

PROPOSITION 5. *Assume* **(A1)**. *If* $[\,]_1 \stackrel{lc}{\equiv} [\,]_2$ *holds, then* $\sum_{X \in \mathcal{A}} E(X) \cdot [X]_1 = \sum_{X \in \mathcal{A}} E(X) \cdot [X]_2$ *holds.* □

We have interests in solving a subproblem of FEMP1, i.e., to find an optimal distribution for FEMP1 among distributions in an equivalence class w.r.t. $\stackrel{lc}{\equiv}$.

Let us consider an equivalence class w.r.t. $\stackrel{lc}{\equiv}$. That is, let us consider a set of distributions $[\,]$ such that $\overline{[e]} = w_e$ for any hyperarc e of G, where w_e's are specified constant reals in $\mathbf{R}_{++}$. We assume here that the constants w_e's satisfy the following condition:

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

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

For a vertex v of G, for convenience, we define $w_v = \sum_{e \in v_{out}} w_e$.

PROPOSITION 6. *Assume* **(A1)** *and* **(A2)** *and that* w_e*'s satisfy* **(C1)** *and* **(C2)**. *For any* $\theta \in \Theta_k$ *and* $v \in V - V_0 - V_f$ *such that* $n_v = k$, $w_v = w_{\theta(v)}$ *holds.* □

Define a distribution $[\,]_+$ of $PT(G)$ as follows:

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

where V_0 and V_f are sets of initial and final vertices of G, respectively. In case that G is clear from the context, we simply write $[\,]_+$ instead of $[\,]_{+,G}$. Furthermore, define a distribution $[\,]_*$ of $\mathcal{A}$ based on the distribution $[\,]_+$ of $PT(G)$ as follows:

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

PROPOSITION 7. *Assume* **(A1)** *and that* w_e*'s satisfy* **(C1)**. *For any* $e \in Eg$ *and* $v \in V - V_f$, *we have:*

$$\sum_{\gamma \in PT(G)\ s.t.\ e \in \gamma} [\gamma]_+ = w_e \quad \textit{and} \quad \sum_{\gamma \in PT(G)\ s.t.\ v \in \gamma} [\gamma]_+ = w_v.$$

□

PROPOSITION 8. *Assume* **(A1)** *and* **(A2)** *and that* w_e*'s satisfy* **(C1)** *and* **(C2)**. *For any* $X \in \mathcal{A}$ *and* $\gamma_1, \gamma_2 \in PT(X)$, $[\gamma_1]_+ = [\gamma_2]_+$ *holds.* □

PROPOSITION 9. *Assume* **(A1)** *and* **(A2)** *and that* w_e*'s satisfy* **(C1)** *and* **(C2)**. *For any vertex* $v \in V - V_f$ *and any hyperarc* e *of* G, *we have* $\overline{[v]}_* = w_v$ *and* $\overline{[e]}_* = w_e$. □

Consider the following minimization problem:

Free Energy Minimization Problem 2 (FEMP2)
minimize :

$$FE_2(P) \stackrel{def}{\equiv} \sum_{X \in \mathcal{A}} [X](\log[X] - 1).$$

subject to :

$$\sum_{\gamma \in PT(G)\,\text{s.t.}\, e \in \gamma} \frac{[X_\gamma]}{n_\gamma} = w_e \qquad \text{(for each hyperarc } e \text{ of } G)$$

$$[X] \geq 0. \qquad (\forall X \in \mathcal{A})$$

Note that in FEMP2 the values $[X]$'s ($X \in \mathcal{A}$) are the variables.

PROPOSITION 10. $FE_2(P)$ *is twice differentiable and convex over the domain* $\mathbf{R}^n_{++}$. □

THEOREM 2. *Assume* **(A1)** *and* **(A2)** *and that w_e's satisfy* **(C1)** *and* **(C2)**. *Furthermore, we assume that $w_e \in \mathbf{R}_{++}$ for every hyperarc e of G and that there exists a strictly feasible point of FEMP2. The distribution $[]_*$ defined by (5) and (6) gives an optimal distribution for FEMP2.*

PROOF. *We first note that by the assumption of this theorem, FEMP2 satisfies Slater's constraint qualification, i.e., there exists a strictly feasible point (i.e., a point satisfying all equality constraints such that $[X] > 0$ for all $X \in \mathcal{A}$). Furthermore, by Proposition 10, the objective function $FE_2(P)$ is twice differentiable and convex over the domain $\mathbf{R}^n_{++}$. Therefore, Karush-Kuhn-Tucker (KKT) conditions provide necessary and sufficient conditions for optimality of FEMP2. KKT conditions of FEMP2 are given by:*

$$\begin{aligned} [X] &\geq 0, & (\forall X \in \mathcal{A}) \quad (7)\\ \sum_{\gamma \in PT(G),\, e \in \gamma} \frac{[X_\gamma]}{n_e} &= w_e & (\forall e \in Eg) \quad (8)\\ \lambda_X &\geq 0, & (\forall X \in \mathcal{A}) \quad (9)\\ -\lambda_X \cdot [X] &= 0, & (\forall X \in \mathcal{A}) \quad (10)\\ \log{[X]} - \lambda_X + \sum_{\gamma \in PT(X)} \sum_{e \in \gamma} \frac{\mu_e}{n_e} &= 0. & (\forall X \in \mathcal{A}) \quad (11) \end{aligned}$$

Define $\tilde{\lambda_X} = 0$ for all $X \in \mathcal{A}$ and

$$\tilde{\mu_e} = \begin{cases} -\log w_e \cdot n_e & \textit{if } t(e) \in V_0, \\ -\log \frac{w_e}{w_{t(e)}} & \textit{otherwise.} \end{cases}$$

Then, it is straightforward to see by Proposition 9 that $[]_*$, $\tilde{\lambda_X}$ $(X \in \mathcal{A})$, and $\tilde{\mu_e}$ $(e \in Eg)$ satisfy KKT conditions. Therefore, $[]_*$ gives an optimal solution of FEMP2. □

Consider the following minimization problem:

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 :

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

Note that the variables of FEMP3 are w_e's $(e \in Eg)$ and recall that w_v's are sums of variables w_e's, i.e., $w_v = \sum_{e \in v_{out}} w_e$. Therefore, the number of variables are reduced from $|\mathcal{A}|$ in FEMP1 (FEMP2) to $|Eg|$ in FEMP3. We often omit the second argument of $FE_3(P, (w_e \mid e \in Eg))$, and simply write $FE_3(P)$ if the context allows.

PROPOSITION 11. *Assume* **(A1)**. *Every minimizer $(w_e \mid e \in Eg)$ of FEMP3 satisfies $w_e > 0$ for each $e \in Eg$.* □

LEMMA 1. *Assume* **(A1)**. *Let us consider a distribution $[]$ and w_e's $(e \in Eg)$ such that $\overline{[e]} = w_e$. Then, we have:*

$$\sum_{X \in \mathcal{A}} \#(x, X) \cdot [X] \quad = \quad \sum_{e \in Eg} \overline{\#}(x, e) \cdot w_e \qquad (\forall x \in \mathcal{M})$$

□

LEMMA 2. *Assume* **(A1)** *and* **(A2)** *and consider constants w_e's which satisfy* **(C1)** *and* **(C2)**. *Consider a distribution $[]_+$ of $PT(G)$ and a distribution $[]_*$ of $\mathcal{A}$ defined by (5) and (6), respectively. Then, we have:*

$$\begin{aligned} &\sum_{X \in \mathcal{A}} E(X) \cdot [X]_* = \sum_{e \in Eg} \overline{E}(e) \cdot w_e, \\ &\sum_{X \in \mathcal{A}} [X]_*(\log[X]_* - 1) = \sum_{e \in Eg} w_e(\log w_e - 1) - \\ &\qquad \sum_{v \in V - V_0 - V_f} w_v(\log w_v - 1) + \sum_{e \in (V_0)_{out}} w_e \cdot \log n_e. \end{aligned}$$

Furthermore, $FE_1(P, []_*) = FE_3(P, (w_e \mid e \in Eg))$ holds. □

As will be shown in Theorem 4, the objective function of FEMP3 is convex. Therefore, FEMP3 have an optimal solution.

THEOREM 3. *Assume* **(A1)** *and* **(A2)**. *Let $(\tilde{w}_e \mid e \in Eg)$ be a minimizer of FEMP3. Then, the distribution $[]_*$ defined by (5) and (6) based on $\tilde{w}_e$'s is a minimizer of FEMP1.*

PROOF. *Let $[]_1$ be a minimizer of FEMP1 and E_1^* be an optimal value of $FE_1(P)$. We know that $[X]_1 > 0$ holds for every $X \in \mathcal{A}$. Let $w_e = \overline{[e]_1}$ for every $e \in Eg$. Consider an optimization problem FEMP2 with these w_e's.*

Note that $w_e > 0$ holds for every $e \in Eg$. By Proposition 2, we have $\overline{[\theta(e)]_1} = \overline{[e]_1}$ for any $e \in Eg$ and any $\theta \in \Theta_k$ with $n_e = k$. Therefore, the values w_e's satisfy the condition **(C2)**. *By Proposition 1, w_e's satisfy the condition* **(C1)**. *Furthermore, $[]_1$ clearly satisfies the constraints of FEMP2. Thus, $[]_1$ gives a strictly feasible point of this FEMP2. So, we can apply Theorem 2.*

Let E_2^* be the optimal value of the objective function of FEMP2. By Theorem 2, a minimizer $[]_2$ of FEMP2 can be given by:

$$\begin{aligned} [\gamma]_+ &= \frac{\prod_{e \in \gamma} w_e}{\prod_{v \in \gamma,\, v \notin V_0,\, v \notin V_f} w_v}, & (12)\\ [X]_2 &= \sum_{\gamma \in PT(X)} [\gamma]_+. & (13) \end{aligned}$$

Since $[]_2$ is a minimizer of FEMP2 and $[]_1$ gives a feasible point of FEMP2, it is clear that $E_1^* \geq E_2^* + \sum_{X \in \mathcal{A}} E(X) \cdot$

$[X]_1$ *holds. Furthermore, we have:*

$$\begin{aligned} E_1^* &\geq E_2^* + \sum_{X\in\mathcal{A}} E(X)\cdot[X]_1 \\ &= E_2^* + \sum_{X\in\mathcal{A}} E(X)\cdot[X]_2 \\ &\qquad \textit{(By } [\,]_1 \overset{lc}{\equiv} [\,]_2 \textit{ and Proposition 5)} \\ &= \sum_{X\in\mathcal{A}} [X]_2(\log[X]_2 - 1) + \sum_{X\in\mathcal{A}} E(X)\cdot[X]_2 \\ &= \sum_{e\in Eg} w_e(\log w_e - 1) \; - \sum_{v\in V,\, v\notin V_0, v\notin V_f} w_v(\log w_v - 1) + \\ &\qquad \sum_{e\in(V_0)_{out}} w_e \log n_e + \sum_{e\in Eg} \overline{E}(e)\cdot w_e. \quad \textit{(By Lemma 2)} \end{aligned}$$

Let F be the last expression of the above transformation.

We have $w_e = \overline{[e]_1}$ ($e \in Eg$) by the definition of w_e's. By Lemma 1,

$$\sum_{e\in Eg} \overline{\#}(x,e)\cdot w_e = \sum_{X\in\mathcal{A}} \#(x,X)\cdot[X]_1 = [x]_0,$$

holds. Recall that w_e's satisfy **(C1)** *and* **(C2)**. *Thus, $(w_e \mid e \in Eg)$ is a feasible point of FEMP3. Let E_3^* be the optimal value of the objective function of FEMP3. Then, $F \geq E_3^*$ holds, which implies $E_1^* \geq E_3^*$.*

On the other hand, let $(\tilde{w_e} \mid e \in Eg)$ be a minimizer of FEMP3. By Proposition 11, $\tilde{w_e} > 0$ holds for every $e \in Eg$. Let $[\,]_*$ be a distribution of $\mathcal{A}$ defined by:

$$\begin{aligned} [\gamma]' &= \frac{\prod_{e\in\gamma} \tilde{w_e}}{\prod_{v\in\gamma,\, v\notin V_0,\, v\notin F} \tilde{w_v}}, \\ [X]_* &= \sum_{\gamma\in PT(X)} [\gamma]', \end{aligned}$$

where $\tilde{w_v} = \sum_{e\in v_{out}} \tilde{w_e}$.

By Proposition 9 and Lemma 1, we can say that $[\,]_*$ satisfies the equality constraints of FEMP1. By $\tilde{w_e} > 0$ ($e \in Eg$), we have $[X]_* > 0$ for every $X \in \mathcal{A}$. Therefore, $[\,]_*$ gives a feasible point of FEMP1. By Lemma 2, $FE_1(P, [\,]_*) = FE_3(P, (\tilde{w_e} \mid e \in Eg)) = E_3^*$. Thus, we have $E_3^* \geq E_1^*$.

In conclusion, we have $E_1^* = E_3^*$, which implies the claim. □

THEOREM 4. *The objective function of FEMP3 is convex over $\mathbf{R}_{++}^m$, where $m = |Eg|$.* □

Therefore, we can solve FEMP1 by solving FEMP3 with a convex programming method([7]). Note that the number of variables are drastically reduced from $|\mathcal{A}|$ to $|Eg|$. The sizes of $|\mathcal{A}|$ and $|Eg|$ are problem-dependent. For example, in a simple 1-dimensional DNA tile assembly system, $|\mathcal{A}|$ is $O(m^n)$ and $|Eg|$ is $O(m^2 n)$, where $m = |\mathcal{M}|$ (number of tiles) and n is the maximum length of tile assemblies.

5. RELATED WORKS

Adleman's work ([2]) on the equilibrium state analysis of linear tile assembly is related to our work. There are two main and important different points between these works: (1) although Adleman's analysis is based on a probabilistic model of chemical reactions, we rely on the concentration based model, and (2) although Adleman's analysis focuses on linear tile assembly alone, we aim to establish a general theory of equilibrium state analysis of HRSs, which can be applied to various HRSs.

Quite recently, Dirks, et al([3]) proposed a method for computing an equilibrium state of interacting RNA molecules based on statistical physics. They nicely extended Mackaskill's partition function computation algorithm for a single RNA molecule to the case of multiple strands, and furthemore succeeded in computing the equilibrium state of interacting RNA molecules by using convex programming after computing partition functions of all strand complexies. The current paper proposes a new method totally different from theirs in the following senses: (1) this paper proposes a general theory for the computation of equilibrium state, (2) although their work assumed the RNA strands interaction in a dilute solution so that the interaction does not happen in an equilibrium state, the current paper assumes the interaction in an equilibrium state among all possible combinations of assemblies and computes exact solutions.

6. CONCLUSIONS

We discussed mathematically on the problem of computing equilibrium state of a combinatorially complex hybridization reaction system in which tremendously many assemblies of molecules are generated from a small number of molecules. In such systems, combinatorial explosion of the number of assemblies makes this problem intractable. In this paper, surprisingly enough, we proposed a novel theoretical framework, in which we can overcome such combinatorial explosion problem when computing equilibrium state. The framework is based on the key concept of *locality* of reaction systems. We formulate this concept using hypergraph theory, and reach to the conclusion that based on some reasonable assumptions, we can efficiently compute the equilibrium state of an HRS by convex programming method.

7. REFERENCES

[1] L. Adleman. Molecular computation of solutions to combinatorial problems. *Science*, 266:1021–1024, 1994.

[2] L. Adleman, Q. Cheng, A. Goel, M. Huang, and H. Wasserman. Linear self-assemblies: Equilibria, entropy, and convergence rates. *unpublished manuscript*, 2000.

[3] R. Dirks, J. Bois, J. Schaeffer, E. Winfree, and N. Pierce. Thermodynamic analysis of interacting nucleic acid strands. *SIAM Review*, 49:65–88, 2007.

[4] G. Gallo, G. Longo, S. Nguyen, and S. Pallottino. Directed hypergraphs and applications. *Discrete Applied Mathematics*, 40:177–201, 1993.

[5] S. Kobayashi. Testing structure freeness of regular sets of biomolecular sequences. In *Preliminary Proceedings of 10th International Meeting on DNA Based Computers*, pages 395–404, 2004.

[6] S. Kobayashi. A new approach to computing equilibrium state of combinatorial chemical reaction systems. Technical Report CS 06-01, Dept. of Computer Science, Univ. of Electro-Communications, 2006.

[7] Y. Nesterov and A. Nemirovskii. *Interior-Point Polynomial Algorithms in Convex Programming.* SIAM, Philadelphia, 1993.