Towards a Calculus of Biomolecular Complexes at Equilibrium

Eric Mjolsness

March 2007

To appear in Briefings in Bioinformatics.

Authors's version.

Abstract

An overview is presented of the construction and use of algebraic partition functions to represent the equilibrium statistical mechanics of multimolecular complexes and their action within a larger regulatory network. Unlike many applications of equilibrium statistical mechanics, multimolecular complexes may operate with various subsets of their components present and connected to the others, the rest remaining in solution. Thus they are variable-structure systems. This aspect of their behavior may be accounted for by the use of "fugacity" variables as a representation within the partition functions.

Four principles are proposed by which the combinatorics of molecular complex construction can be reflected in the construction of their partition functions. The corresponding algebraic operations on partition functions are multiplication, addition, function composition, and a less commonly used operation called contraction. Each has a natural interpretation in terms of probability distributions on multimolecular structures. Possible generalizations to nonequilibrium statistical mechanics are briefly discussed.

1 Introduction

With the recent flourishing of systems biology, interest has increased in methods for quantitatively modeling the effects that biomolecular complexes have on the larger regulatory networks and regulatory systems within which they are embedded. For example, transcription complexes can integrate a number of transcription factor inputs into an overall decision for or against the initiation of transcription of a specific gene. Thus it provides a probability per unit time, or a rate, of transcription initiation. How can we quantitatively model this kind of "integration" of inputs? In the transcription complex case we would like to translate from structural aspects of molecular interactions within the complex, such as protein:DNA binding and protein:protein interactions, which may be partly known and partly hypothesized, to some kind of rate law for the process of transcription. More generally we would like to translate from structure and experimentally measurable parameters to an algebraic formula or very fast algorithm, depending on the measurable parameters, for evaluating a rate law governing a process mediated by the complex. From such a rate law we can create an ordinary differential equation model for the effect of the complex within a larger network. If we also receive an algebraic formula for the standard deviation of actual rate from expected rate, as is possible in statistical mechanics, then we can create more realistic stochastic models such as stochastic differential equations.

Quite a variety of techniques can be used to achieve the goal of constructing such quantitative models. They can be divided into equilibrium and nonequilibrium models for the biomolecular complex. Either can be used within a network-scale model, which must generally be nonequilibrium in a living system. Equilibration can happen at a fast time scale for the complex, which forms a small part of a larger system that changes on a slow time scale and is at equilibrium; the result is a quasi-equilibrium model of the complex and a rate law for its effect on the larger system.

In this note I will review a simplified way of thinking about these problems that can straightforwardly yield algebraic rate laws for the activities of multimolecular complexes in quasi-equilibrium, starting from hypotheses about the interaction connectivity of their state variables having to do with conformation and binding sites. Foundations of equilibrium and nonequilibrium statistical mechanics are recapitulated and related to one another in Section 2.1. Very small examples equilibrium models are discussed in Section 2.2. More complex algebraic models and partition functions can then be built up step by step, using four main principles of construction introduced in Section 2.3. Multiplication of partition functions corresponds to independent probability distributions. Addition corresponds to mixtures of probability distributions. Function composition corresponds to a tree-like (acyclic) topology of interactions between subcomplexes that may or may not bind to one another. Cycles can be added to such a topology using an "contraction" operation on a partition function. Applications of each of these operations will be shown in the examples of Section 2.4 and Section 3, drawn largely from quasiequilibrium models of transcriptional regulation and allosteric enzymes.

2 Theory

2.1 Equilibrium and Nonequilibrium Statistical Mechanics

This section serves to encapsulate the needed elementary statistical mechanics in simple notation.

In an *equilibrium* statistical mechanics model, the probability of any discrete state I of the multimolecular complex, including all information about its discrete conformation and binding status, is proportional to the Boltzmann factor $\exp(-\beta G_I)$. Here G_I is the Gibbs free energy of the state and β is inversely proportional to the temperature: $\beta = 1/(k\,T)$ where k is Boltzmann's constant in appropriate units. Consequently the relative probabilities of any two states I and J have the ratio

$$p_I / p_J = \exp(-\beta\,(G_I - G_J))\,. \tag{1}$$

The normalization factor for the probabilities p_I is a very important function, the "partition function" $Z(\beta)$, that can be constructed from G_I:

$$Z(\beta) = \sum_I \exp(-\beta\,G_I) \quad \Rightarrow \quad p_I = \exp(-\beta\,G_I)/Z(\beta) \tag{2}$$

In a *nonequilibrium* model, by contrast, we postulate nonnegative transition rates (transition probabilities per unit time) K_{IJ} from state J to state I, most of which are zero. One can imagine the possible transitions as "reactions" of the form

$$p_J \underset{K_{JI}}{\overset{K_{IJ}}{\rightleftharpoons}} p_I \tag{3}$$

which conserve total probability $\sum_I p_I = 1$. The discrete-state probabilities p_I then must evolve exactly according to the "master equation" that would ensue from the law of mass action for such a set of reactions:

$$\frac{d\, p_I(t)}{d\, t} = \sum_J K_{IJ}\, p_J(t) - \left(\sum_J K_{JI}\right) p_I(t) \equiv \sum_J H_{IJ}\, p_J(t) \ . \tag{4}$$

Here H is a matrix whose columns each add up to zero.

Equilibrium and nonequilibrium formulations of statistical mechanics are related, in that equilibrium can be achieved by constraining the forward and backward rates K_{IJ} and K_{JI} so as to satisfy the condition of "detailed balance" ($K_{IJ}\, p_J^* = K_{JI}\, p_I^*$ for some p_I^*). This implies a steady state for a Boltzmann distribution defined by p_I^* :

$$\begin{aligned} K_{IJ} / K_{JI} = p_I^* / p_J^* \quad &\equiv \exp\, -\beta\,(G_I - G_J) \\ \Rightarrow \quad \sum_J (K_{IJ}\, p_J^* - K_{JI}\, p_I^*) &= 0 \\ \Rightarrow \quad d\, p_I^* \ / d\, t \ &= 0 \quad \text{(steady state).} \end{aligned}$$

The condition $K_{IJ} / K_{JI} = p_I^* / p_J^*$ is satisfiable by some probabilities p_I^*, if and only if the product of rates K_{IJ} around any cycle is the same whether traversing the cycle forwards or backwards. In this case, equilibrium statistical mechanics arises as the infinite-time endpoint of nonequilibrium statistical mechanics. Further discussion of this point is in the Supplemental Material Section 5.1.

The nonequilibrium approach may be essential for many molecular machines, possibly in combination with equilibrium models for selected substructures. A general and systematic approach to constructing nonequilibrium stochastic models taking a form similar to the H matrix in Equation 4, starting from models specified in terms of reactions similar to Equation 3 augmented with parameters for the reactants, and resulting in simulation algorithms, has been proposed [1].

I now turn to the algebraic construction of solvable free energy functions G_I .

2.2 Equilibrium for binding sites

2.2.1 A single binding site

As an example, consider a single binding site which may be unoccupied or may be occupied by a single molecule of only one particular molecular species A. The change in free energy for occupying the binding site is the sum of the binding energy (which we expect to be negative) and the change in free energy due to taking one molecule of A out of solution:

$$\Delta\, G_{\text{occupied} \leftarrow \text{unoccupied}} = \Delta\, G_{\text{binding}} - k\, T \log z_A$$

where the "fugacity" z_A is proportional to concentration of A in a dilute solution. Then by Equation 1

$$p_{\text{occupied}} / p_{\text{unoccupied}} = e^{-\beta \Delta G_{\text{binding}}} z_A \equiv \omega_A z_A .$$

The partition function of Equation 2 provides the normalization for the two relative probabilities $\omega_A z_A$ and 1:

$$Z(z_A) = 1 + \omega_A z_A$$

and the probability of occupancy is

$$p_{\text{occupied}} = \frac{\omega_A z_A}{1 + \omega_A z_A} = \frac{\partial \log Z(z_A)}{\partial \log z_A} .$$

The logarithmic derivative of $Z(z_A)$ is typical of how Z can be used to calculate meaningful averages. A graph representation of this probability distribution would consist of a single binary variable $s \in \{0, 1\}$, denoted by an open circle, in total isolation from any other variables (Figure 1a). The value of s is the number of occupying A's, either zero or one.

2.2.2 Two independent sites

To build up larger models we can use rules for appropriately combining partition functions. For two binding sites that have no influence on one another, and are thus independent, the partition functions multiply:

$$Z(z_1, z_2) = (1 + \omega_1 z_1)(1 + \omega_2 z_2)$$

so that for example

$$p_{1\,\text{occupied}} = \frac{\partial \log Z(z_1, z_2)}{\partial \log z_1} = \frac{\omega_1 z_1}{1 + \omega_2 z_2}$$

as expected. The monomial terms in the expanded form of Z,

$$Z(z_1, z_2) = 1 + \omega_1 z_1 + \omega_2 z_2 + \omega_1 \omega_2 z_1 z_2,$$

correspond to the four different possible binding states and their probabilities.

On the other hand if sites 1 and 2 both bind A and we want to know the average total number of A molecules bound, the answer given by a different logarithmic derivative of Z (Supplemental Material Section 5.2).

The graphical model for this system would consist of two binary variable nodes $s_i \in \{0, 1\}$, one for each site, with no connecting link between them or any other variable (Figure 1b).

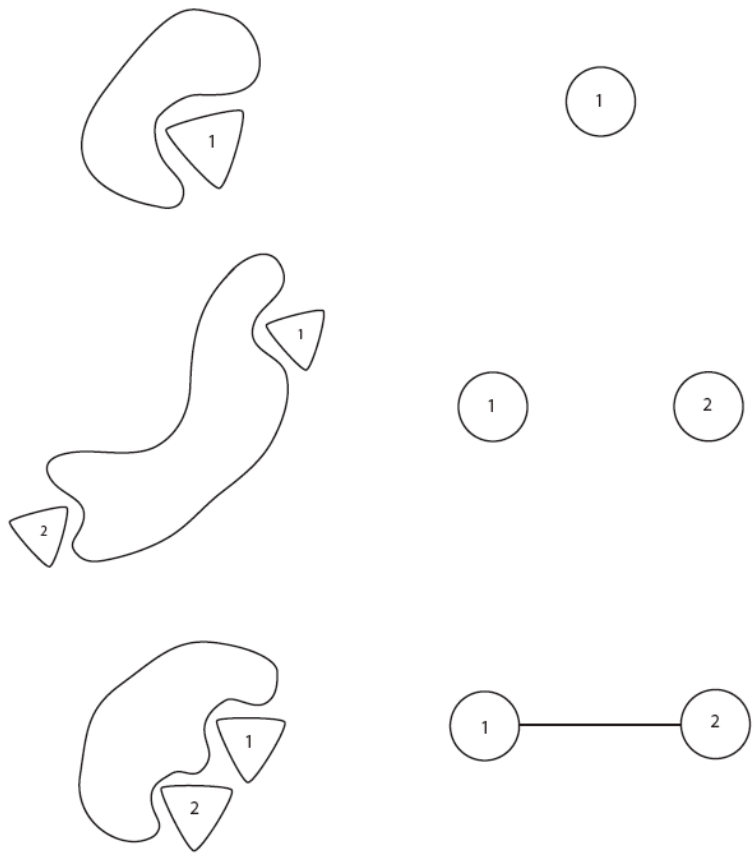

Figure 1. Illustration of elementary partition functions in dilute solution. Rows: a,b,c. (a) Left: single binding site, occupied. Right: Single binary occupancy variable uncoupled to all others, representing this situation. Partition function is $1 + \omega_1 z_1$. (b) Left: Two independent binding sites, both occupied. Right: Two binary occupancy variables representing this situation, uncoupled. Partition function is $Z(z_1, z_2) = (1 + \omega_1 z_1)(1 + \omega_2 z_2)$. (c) Left: Two binding sites with energetic interaction (synergistic or antisynergistic) between the two occupying molecules. Right: Two binary occupancy variables representing this situation, energetically coupled. Partition function is $1 + \omega_1 z_1 + \omega_2 z_2 + \omega_1 \omega_2 \omega_{1\,2} z_1 z_2$.

2.2.3 Two non-independent sites

If we now allow energetic interactions between two binding sites $b = 1$ and $b = 2$ that can each be empty or occupied by molecules of species 1 or 2 respectively, and no other internal states, then

$$Z(z_1, z_2) = \sum_{\{s \,|\, s_i \in \{0,1\}\}} z_1{}^{s_1} \, z_2{}^{s_2} \, \omega_1{}^{s_1} \, \omega_2{}^{s_2} \, \omega_{1\,2}{}^{s_1 \, s_2} = 1 + \omega_1 z_1 + \omega_2 z_2 + \omega_1 \omega_2 \omega_{1\,2} z_1 z_2$$

This situation is illustrated in Figure 1c.

A protein with a single binding site that can be empty or occupied by species 1 or 2 would be modeled the same way except that both occupiers cannot be present simultaneously, i.e. $\overline{s_1 \wedge s_2}$ must be true, hence the impossible state $s_1 = s_2 = 1$ is omitted from the sum over all possible states in the partition function and $Z(z_1, z_2) = 1 + \omega_1 z_1 + \omega_2 z_2$:

$$Z(z_1, z_2) = \sum_{\{s \,|\, s_i \in \{0,1\} \wedge \overline{s_1 \wedge s_2}\}} z_1{}^{s_1} z_2{}^{s_2} \omega_1{}^{s_1} \omega_2{}^{s_2} \omega_{1\,2}{}^{s_1 s_2} = 1 + \omega_1 z_1 + \omega_2 z_2$$

The generalization to Boltzmann distributions in general is discussed in the Supplemental Material, Section 5.3.

If the protein is itself regarded as another species that can be present or absent, with fugacity z_0, then it must be present, so $s_0 \wedge \overline{s_1 \wedge s_2}$ must be true, and the partition function is $Z(z_1, z_2) = z_0 (1 + \omega_1 z_1 + \omega_2 z_2)$. Likewise, a heterodimer consisting only of species 1 and 2 with no internal states would satisfy $s_1 \wedge s_2$ and therefore $Z(z_1, z_2) = \omega_{1\,2} z_1 z_2$. (This is a trivial case since there is only one state, but it will be useful when the species are given internal states as well.) In each case, as for any probability generating function, the coefficients can be normalized to give the probabilities of each possible configuration of bindings.

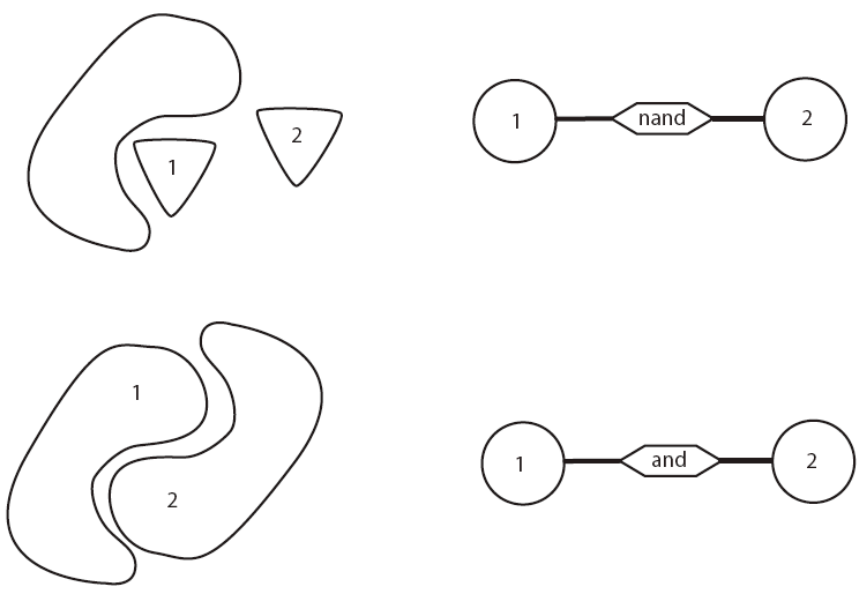

Figure 2. Illustration of elementary partition functions in dilute solution. (a) Left: A protein with a single binding site that can be empty or occupied by species 1 or 2, but not occupied by both. $1 + \omega_1 z_1 + \omega_2 z_2$. Right: logical coupling $\overline{s_1 \wedge s_2}$ between two binary occupancy variables representing this situation. (b) Left: A heterodimer of two molecular species "bound" to a fictitious enclosing complex. $Z(z_1, z_2) = \omega_{1\,2} z_1 z_2$. Right: logical coupling $s_1 \wedge s_2$ between two binary complex-membership variables representing this situation.

Thus partition functions may be expressed as polynomials in fugacity variables. This is a particularly convenient notation for molecules in a dilute solution which acts as a reservoir, since in that case fugacities z_i are proportional to concentrations $c_i = [S_i]$. Such polynomial partition functions can be put into a form with homogeneous degree by introducing the complementary fugacity variables z_i^+ and z_i^- and substituting $z_i = z_i^+ / z_i^-$: $Z^{\text{homog}}(\boldsymbol{z}^+, \boldsymbol{z}^- \mid \omega) = Z(\boldsymbol{z}^+ / \boldsymbol{z}^- \mid \omega)(\prod_i z_i^-)$. No information is lost since $Z(\boldsymbol{z} \mid \omega) = Z^{\text{homog}}(\boldsymbol{z}^+ = z, \boldsymbol{z}^- = \mathbf{1} \mid \omega)$.

2.3 Principles for combining partition functions

There is a useful set of principles for creating partition functions that reflect the structure of multimolecular complexes. It can't hurt to name the principles after the algebraic operations on partition functions that they invoke: *multiplication, addition, composition,* and *contraction.*

Multiplication. From the foregoing example of two independent binding sites, a first principle for combining partition functions is: The product of two partition functions corresponding to probability distributions $P_1(s_1)$ and $P_2(s_2)$ is itself a partition function representing the independent distribution $P_1(s_1)\,P_2(s_2)$. In other words, *the partition functions of independent random variables multiply.* In this case the logarithmic derivatives just add up.

Addition. A second principle results from adding partition functions. The weighted sum $w_1\, Z_1(\boldsymbol{z}) + w_2\, Z_2(\boldsymbol{z})$ reweights all the monomial terms within either $Z_1(\boldsymbol{z})$ or $Z_2(\boldsymbol{z})$. Thus, the nonnegatively *weighted sum of partition functions represents a mixture distribution* of the component probability distributions with related weights, in this case $w'_1\, P_1(\boldsymbol{s}) + w'_2\, P_2(\boldsymbol{s})$.

Composition. A third principle is a powerful tool for constructing complex structures and will be illustrated in Section 2.4: that *functional composition of partition functions corresponds to a tree-like structure* in the construction of the molecular system represented. The functional composition is acheived by substituting a whole partition function Z_f for an appropriate fugacity variable z_i occurring within another partition function Z_c, obtaining $Z_c(Z_f(\boldsymbol{z}'), \boldsymbol{z})$. The assumed molecular tree structure is reflected in the tree structure of partition functions and their arguments. Logarithmic derivatives may then be taken using the chain rule of differential calculus, as is done in Equation 6 in Section 3.1 below. A similar phenomenon arises in the theory of birth and death processes or branching processes, in which a single generating function may be composed with itself many times [2].

Contraction. A fourth principle is not quite as simple. A special *contraction* operation on partition functions, defined in Section 3.3, allows cycles to be introduced into the treelike, noncyclic structure resulting from the operation of the third principle. This is important due to the many non-treelike structures, such as rings, that may be present in multimolecular complexes. The resulting computations appear to be generally more difficult, so that a premium is placed on minimizing the number of contractions.

These principles can be applied repeatedly in various orders to generate interesting structure, as we will show in the remainder of the paper using their italicized names to highlight how each principle is employed. Together, these principles point towards a symbolic "calculus" for creating and reasoning about equilibrium multimolecular complex models.

2.4 Branching tree structure

Dyson [3] first solved the equilibrium statistical mechanics of a model with branching structure. Here we will provide notation to make this type of calculation convenient for molecular complexes that can be built up in solution.

2.4.1 Transcriptional regulation by multiple binding sites

In a simple quasi-equilibrium model of transcriptional regulation of gene i, which has a set of transcription factor binding sites labelled by (i, b), by transcription factors j, it is assumed that the complex has two states that permit or prohibit transcriptional initiation. We start out using the *addition principle* just as for the single-binding-site example, except that now the single binary variable $s_i = \pm 1$ refers to conformation and not occupancy. We give these two alternatives fugacity-like variables $z_i^{\pm}$, creating a homogeneous polynomial in $\boldsymbol{z}$:

$$Z_i\,(\boldsymbol{z}) = \omega_i\, z_i^{+} + z_i^{-}$$

If we freeze the conformation variable s_i, then the transcription factor binding sites are assumed to be occupied independently. Here we simplify and assume each site (i, b) is specific for a single transcription factor $j(i, b)$, so that it has partition function

$$\Xi_{(i\,b)}^{(s_i)} = 1 + \omega_{(i\,b)}^{(s_i)}\, z_{j(i\,b)}\,.$$

By the *multiplication principle*, the partition functions $\Xi_{(i\,b)}^{(s_i)}$ multiply up over the sites b yielding $\prod_{b=1}^{B} \Xi_{(i\,b)}^{(s_i)}$. By the *composition principle*, we can substitute this product for the fugacity $z_i^{(s_i)}$ in $Z_i\,(\boldsymbol{z})$. The resulting model is

$$Z_i\,(\zeta_i, \boldsymbol{z}) = \zeta_i\, \omega_i \prod_{b=1}^{B} \left(1 + \omega_{(i\,b)}^{+}\, z_{j(i\,b)}\right) + \prod_{b=1}^{B} \left(1 + \omega_{(i\,b)}^{-}\, z_{j(i\,b)}\right)$$

where we have introduced temporarily the extra fugacity-like variable ζ_i to mark terms associated with $s_i = 1$, i.e. the transcriptionally active states. The fraction of maximal transcriptional activation is then [4]

$$\left\langle \frac{s_i + 1}{2} \right\rangle = \frac{\partial \log Z_i}{\partial \log \zeta_i}\bigg|_{\zeta_i = 1} = \frac{\omega_i \prod_{b=1}^{B} \left(1 + \omega_{(i\,b)}^{+}\, z_{j(i\,b)}\right)}{\omega_i \prod_{b=1}^{B} \left(1 + \omega_{(i\,b)}^{+}\, z_{j(i\,b)}\right) + \prod_{b=1}^{B} \left(1 + \omega_{(i\,b)}^{-}\, z_{j(i\,b)}\right)} \tag{5}$$

This model, inspired by the classic Monod-Wyman-Changeux (MWC) model for allosteric enzymes [5], can serve as a foundation for deriving [4] artificial neural net (ANN) models of transcriptional regulation that have been used in models of *Drosophila* development (e.g. [6]). If factor j has just one binding site regulating gene i, the activation as a function of z_j is an $n = 1$ Hill function with no cooperativity. But multiple binding sites for a single factor j can add cooperativity to the transcriptional response of gene i to factor j.

The interaction graph corresponding to this model consists of a parent node s_i connected separately to each of a set of children nodes $s_{i\,b} \in \{0, 1\}$ representing the occupancy of the corresponding sites. There are no direct connections among the children. This situation is is illustrated in Figure 3a.

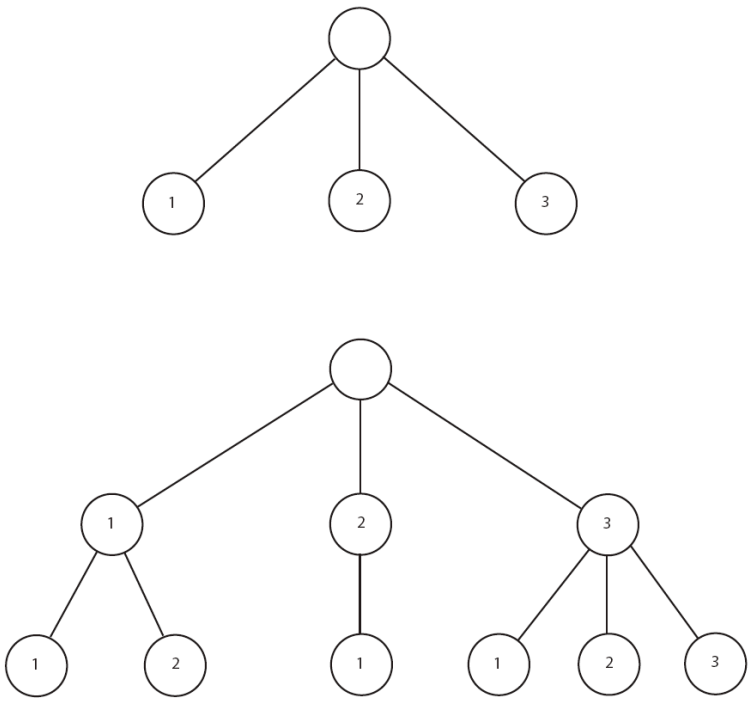

Figure 3. (a) Parent node energetically coupled to children nodes, which are not energetically coupled amongth themselves. No lateral interactions are present. (b) Two layers of tree coupling among three layers of tree nodes. No lateral interactions are present.

Further partition function engineering of this general flavor was performed in order to generalize the MWC model to allosteric enzymes with multiple activators, inhibitors, and/or substrates [7]. Logarithmic derivatives were used to derive rate laws for the action of such enzymes for inclusion within a larger-scale ordinary differential equation model for the synthesis of branched chain amino acids in *E. coli* .

3 Examples

3.1 Transcriptional Regulation by Hierarchical Cooperative Activation

We may take the foregoing model one step further by interposing an extra level of hierarchical structure in transcriptional activation corresponding to *modules* of interacting binding sites. Using *addition, multiplication*, and *composition* exactly as before,

$$Z_i\,(\zeta_i\,,\,\boldsymbol{z}) = \zeta_i\;\omega_i\;\prod_{m=1}^{M(i)} Z_{i\,m}^{+} + \prod_{m=1}^{M(i)} Z_{i\,m}^{-}$$

$$Z_{i\,m}^{(s_i)}\left(\eta_{i\,m}\,,\,\boldsymbol{z}\right) = \eta_{i\,m}\;\omega_{i\,m}^{(s_{i\,m})}\;\prod_{b=1}^{B(i,m)} \Xi_{i\,m\,b}^{+} + \prod_{b=1}^{B(i,m)} \Xi_{i\,m\,b}^{-}$$

$$\Xi_{i\,m\,b}^{(s_{i\,m})} = 1 + \omega_{i\,m\,b}^{(s_{i\,m\,b})}\;z_{j(i\,m\,b)}$$

This gives us the Hierarchical Cooperative Activation (HCA) model of transcriptional regulation. Its graph consists of a parent node, the transcriptional activation variable s_i , connected or linked to its children nodes, the module activation variables $s_{i\,m}$ (none connected to each other), each of which is linked to grandchildren node occupancy variables $s_{i\,m\,b}$ (Figure 3b). By letting the index i vary in these equations, we may describe an entire network of mutually regulating transcription factors.

An example calculation using logarithmic derivatives and the chain rule is to calculate the fractional site occupancy of one binding site, which may be an observable quantity:

$$f_{(i\,m\,b)} = \sum_{s_i=0,1} \Pr(s_i) \sum_{s_{i\,m}=0,1} \Pr(s_{i\,m} \mid s_i)\,\Pr(s_{i\,m\,b} \mid s_{i\,m})\,, \tag{6}$$

where each Pr can be further calculated as in Equation 5. The brief calculation is shown in the Supplemental material, Section 5.4 . It provides an example of calculating with a “biomolecular calculus”.

3.2 1D Chain

An example of a different sort has state information, but no occupancy information and no variable connectivity due to varying occupancy of binding sites. Instead there is one binary variable $s_i \in \{\pm 1\}$ at every node in a one-dimensional chain of N identical nodes. We may regard the first node as the “top” one in a downward growing lineage tree that never branches, having *additive* partition function at the top level:

$$Z_1\,(\boldsymbol{z}) = \omega^{+}\;z_1^{+} + \omega^{-}\;z_1^{-} = (\,1\quad 1\,)\cdot\begin{pmatrix}\omega^{+} & 0\\ 0 & \omega^{-}\end{pmatrix}\cdot\begin{pmatrix}z_1^{+}\\ z_1^{-}\end{pmatrix}. \tag{7}$$

Then the *composition principle* gives the first substitution to make in Equation 7, to reach the second level:

$$z_1^{(s)} \mapsto \omega^{+}\;\omega^{s+}\;z_2^{+} + \omega^{-}\;\omega^{s-}\;z_2^{-}\;.$$

Similarly at depth i in the trivial nonbranching tree (i.e. at position i in the chain), the *composition principle* can be written

$$\begin{pmatrix}z_i^{+}\\ z_i^{-}\end{pmatrix} \mapsto \begin{pmatrix}\omega^{++} & \omega^{+-}\\ \omega^{-+} & \omega^{--}\end{pmatrix}\cdot\begin{pmatrix}\omega^{+} & 0\\ 0 & \omega^{-}\end{pmatrix}\cdot\begin{pmatrix}z_{i+1}^{+}\\ z_{i+1}^{-}\end{pmatrix}\,.$$

This rule is easy to apply recursively to Equation 7, resulting in the usual solution for a 1D Ising model with chain topology as detailed in the Supplemental Material, Section 5.5 .

Similar chain models, albeit nonhomogeneous, can be used to augment the HCA model by describing the equilibrium of a chain of transcription factor binding sites that compete for occupancy by overlapping with one another along DNA [4].

3.3 Contraction example: 1D Ring

A final example demonstrates the *contraction principle* by which we can go beyond treelike topologies by adding cycles. The graph interpretation of contraction is to identify two nodes representing state variables s_i and s_j, demanding that they are identically equal and that any redundant probability factors be removed. Graphically this operation can add a cycle to a tree. Here we will demonstrate it on the trivial tree consisting of the chain in the 1D Ising model of length $N+1$, tying the first and last state variables to each other as $s_{N+1} = s_1$.

In order to "remember" the value of s_1 we introduce the temporary fugacity-like variable ξ in the previous calculation:

$$Z_1(\boldsymbol{z}) = \omega^+ \; z_1^+ \, \xi + \omega^- \; z_1^-$$

The last node must correct for the overcounting of $\omega^{(s_1)}$ and must cancel out ξ iff $s_{N+1} = s_1$:

$$\begin{pmatrix} z_{N+1}^+ \\ z_{N+1}^- \end{pmatrix} \mapsto \begin{pmatrix} \xi^{-1} / \omega^+ \\ 1 / \omega^- \end{pmatrix}$$

Then we pick off the coefficient of the ξ^p term of the series expansion in of $Z_N(\xi, \boldsymbol{z})$ in ξ, denoted here $\mathrm{Coef}[Z_N(\xi, \boldsymbol{z}), \xi, p]$, for $p = 0$, in this way performing the *contraction* operation:

$$Z_N(\boldsymbol{z}) \,|_{z_{N+1} \mapsto z_1} = \mathrm{Coef}\left[\begin{pmatrix} \xi & 1 \end{pmatrix} \cdot \ldots \cdot \begin{pmatrix} \xi^{-1} / \omega^+ \\ 1 / \omega^- \end{pmatrix}, \xi, 0\right] = \mathrm{Trace}\left[\Omega^N\right] \tag{8}$$

This is the usual solution for a homogeneous ring (known in statistical mechanics as the one-dimensional homogeneous Ising model with periodic boundary conditions) as detailed in the Supplemental Material, Section 5.5.

For example, for $N = 4$ the resulting partition function is

$$Z(J, \mu) = 4 + 2\, e^{-4 J \beta} + e^{4 \beta (J-\mu)} + 4\, e^{-2 \beta \mu} + 4\, e^{2 \beta \mu} + e^{4 \beta (J+\mu)}$$

and for $N = 6$ it is

$$Z(J, \mu) = 2\, e^{-6 J \beta} + 12\, e^{-2 J \beta} + 6\, e^{2 J \beta} + e^{6 J \beta - 6 \beta \mu} + 6\, e^{2 J \beta - 4 \beta \mu} + 9\, e^{-2 J \beta - 2 \beta \mu} + 6\, e^{2 J \beta - 2 \beta \mu} + 9\, e^{-2 J \beta + 2 \beta \mu} + 6\, e^{2 J \beta + 2 \beta \mu} + 6\, e^{2 J \beta + 4 \beta \mu} + e^{6 J \beta + 6 \beta \mu}.$$

A general notation and integral expression for the contraction operation is given in the Supplemental Material, Section 5.6.

4 Discussion

The foregoing examples can be augmented with many others by repeatedly applying the four principles proposed for constructing partition functions. For example, one could create a ring of trees topology. Each of the principles has an interpretation in terms of interaction graphs in equilibrium statistical mechanics, now widely applied as "Markov random fields" in pattern recognition or "graphical models" in machine learning. Multiplication corresponds to disconnected subgraphs. Addition corresponds to a mixture model gated by a discrete selection variable. Composition corresponds to a tree topology, which unlike conventional graphical models may have a variable structure due to subtrees that can be present or absent. Contraction corresponds to identifying two existing nodes and thereby possibly creating cycles. These latter two principles still require experience to apply correctly since they are not yet fully formalized and automated using a computer algebra representation such as that of [8].

5 Supplemental Material: Some Mathematical Details

5.1 Relation of equilibrium and nonequilibrium statistical mechanics

The condition $K_{IJ} / K_{JI} = p_I^* / p_J^*$ is satisfiable by some probabilities p_I^*, if and only if the product of rates K_{IJ} around any cycle is the same whether traversing the cycle forwards or backwards. In this case, equilibrium statistical mechanics arises as the infinite-time endpoint of nonequilibrium statistical mechanics.

What if K doesn't satisify the cycle condition? Due fundamentally to energy conservation, K may nonetheless represent an approximation to some underlying $\tilde{K}$ which does satisfy the cycle condition and does have an equilibrium solution $\tilde{p}^*$. Before finally going to equilibrium, $\tilde{p}(t)$ could spend a long time at a "quasi-stationary state" of $\tilde{K}$, in the neighborhood of a steady state p^* of K that doesn't satisfy detailed balance. The timescale for this penultimate behavior can be revealed by studying the least negative eigenvalues $\tilde{\lambda}$ of the eigenvalue problem $\tilde{H}\, p = \tilde{\lambda}\, p$ arising from Equation 4 for $\tilde{K}$.

There are prospects for generalizing to nonequilibrium statistical mechanics, though that will be more difficult. We have provided ways to calculate the relative probabilities ω_I of states indexed by I, where in a fundamental description of a biochemical system

$$K_{IJ} / K_{JI} = \omega_I / \omega_J$$

However nonequilibrium statistical mechanics requires a lot of extra rate information such as the sparse matrix

$$\rho_{IJ} = \sqrt{K_{IJ}\, K_{JI}}$$

in order to determine the K's. Then the equilibrium state is an eigenvector of K with zero eigenvalue. For nonequilibrium, the eigenvectors corresponding to the least negative eigenvalues of K will be of most

interest. How these may be related to the graph of interactions between molecular state variables, including site occupancies, is a subject for future work. We may expect that inequalities and approximations will be easier to come by than exact results.

A general model specification and simulation framework, defined in terms of the matrix H of Equation 4, is formalized in [1].

5.2 Two binding site calculations

If sites 1 and 2 both bind A and we want to know the average total number of A molecules bound, the answer is the logarithmic derivative

$$\langle \text{bound } A \rangle = \frac{\partial \log Z(z_1, z_2)\,|_{z_1, z_2 = z_A}}{\partial \log z_A} = \frac{\partial \log Z(z_A, z_A)}{\partial \log z_A} = \frac{\omega_1 z_A}{1 + \omega_1 z_A} + \frac{\omega_2 z_A}{1 + \omega_2 z_A},$$

again as expected. If the two sites are identical the calculation simplifies:

$$Z(z_1, z_2) = (1 + \omega z_1)(1 + \omega z_2)$$

$$\langle \text{bound } A \rangle = \frac{\partial \log (1 + \omega z_A)^2}{\partial \log z_A} = 2 \frac{\omega z_A}{1 + \omega z_A}.$$

5.3 Relation to Boltzmann distributions

The partition functions we have constructed all take the general form

$$Z(\boldsymbol{z} \mid \omega) = \sum_{\{\boldsymbol{s} \mid P(\boldsymbol{s})\}} \left(\prod_i z_i^{s_i} \, \omega_i^{s_i} \right) \prod_{\{i\,j \text{ interacting}\}} \omega_{i\,j}^{s_i\, s_j}$$

though higher-order interactions are also possible. Here $P(\boldsymbol{s})$ is a predicate on the logically allowed combinations of state variables $\boldsymbol{s}$. This is equivalent to the usual formulation

$$Z(\boldsymbol{z} \mid J) = \sum_{\{\boldsymbol{s} \mid s_i \in \{0,1\}\}} \exp(-\beta E), \quad \text{where}$$

$$E = \sum_i \mu_i s_i + \sum_{i\,j} J_{i\,j} s_i s_j + \ldots$$

with

$$\exp(-\beta \mu_i) = \exp(-\beta G_i) = z_i \omega_i$$
$$\exp(-\beta J_{i\,j}) = \exp(-\beta G_{i\,j}) = \omega_{i\,j}$$

We change notation from energy E to free energy G because of the presence in the surrounding solution of a reservoir of possible binding site occupants.

5.4 Binding site occupancy calculation

An example calculation using the chain rule with logarithmic derivatives is to calculate the fractional site occupancy of one binding site, which may be an observable quantity [4]:

$$\begin{aligned} f_{(i\,m\,b)} &= \frac{\partial \log Z_i(\zeta_i)}{\partial \log z_{j(i\,m\,b)}} \Bigg|_{\zeta_i=\eta_m=1} \\ &= \sum_{s_i} \frac{\partial \log Z_i(\zeta_i)}{\partial \log Z_{i\,m}^{(s_i)}} \sum_{s_{i\,m}} \frac{\partial \log Z_{i\,m}^{(s_i)}}{\partial \log \Xi_{i\,m\,b}^{(s_{i\,m})}} \frac{\partial \log \Xi_{(i\,m\,b)}^{(s_{i\,m})}}{\partial \log z_{j(i\,m\,b)}} \\ &= \sum_{s_i=0,1} \Pr(s_i) \sum_{s_{i\,m}=0,1} \Pr(s_{i\,m} \mid s_i) \Pr(s_{i\,m\,b} \mid s_{i\,m}) , \end{aligned} \tag{9}$$

where each Pr can be further calculated as in Equation 5. This is an example of calculating with a "biomolecular calculus".

5.5 1D Chain and ring calculations

Ising model calculation details for the chain:

The two states $s_1 = \pm 1$ can be assigned relative probabilities determined by

$$\omega^{(s_1)} = e^{\beta \mu s_1} .$$

Note $\omega^+ \omega^- = 1$. Let the interaction energy between the parent and child state variables be

$$\omega^{(s_1 s_2)} = e^{\beta J s_1 s_2} .$$

This rule is easy to apply recursively to Equation 7, resulting in

$$Z_n(\boldsymbol{z}) = \begin{pmatrix} 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} \omega^+ & 0 \\ 0 & \omega^- \end{pmatrix} \cdot \left[\begin{pmatrix} \omega^{++} & \omega^{+-} \\ \omega^{-+} & \omega^{--} \end{pmatrix} \cdot \begin{pmatrix} \omega^+ & 0 \\ 0 & \omega^- \end{pmatrix} \right]^{n-1} \cdot \begin{pmatrix} z_n^+ \\ z_n^- \end{pmatrix}$$

which can be reexpressed as

$$Z_N(\mathbf{1}) = \begin{pmatrix} 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} \sqrt{\omega^+} & 0 \\ 0 & \sqrt{\omega^-} \end{pmatrix} \cdot \left[\begin{pmatrix} \omega^{++}\,\omega^+ & \omega^{+-} \\ \omega^{-+} & \omega^{--}\,\omega^- \end{pmatrix} \right]^{N-1} \cdot \begin{pmatrix} \sqrt{\omega^+} & 0 \\ 0 & \sqrt{\omega^-} \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

This can be evaluated explicitly by diagonalizing the central matrix

$$\Omega \equiv \begin{pmatrix} \omega^{++}\,\omega^+ & \omega^{+-} \\ \omega^{-+} & \omega^{--}\,\omega^- \end{pmatrix} = \begin{pmatrix} e^{\beta(J+\mu)} & e^{-\beta J} \\ e^{-\beta J} & e^{\beta(J-\mu)} \end{pmatrix} .$$

Ising model calculation details for a ring:

In order to "remember" the value of s_1 we introduce the temporary fugacity-like variable ξ in the previous calculation:

$$Z_n(\xi, \boldsymbol{z}) = (\xi \quad 1) \cdot \begin{pmatrix} \omega^+ & 0 \\ 0 & \omega^- \end{pmatrix} \cdot \left[\begin{pmatrix} \omega^{++} & \omega^{+-} \\ \omega^{-+} & \omega^{--} \end{pmatrix} \cdot \begin{pmatrix} \omega^+ & 0 \\ 0 & \omega^- \end{pmatrix} \right]^{n-1} \cdot \begin{pmatrix} z_n^+ \\ z_n^- \end{pmatrix}.$$

The last node must correct for the overcounting of $\omega^{(s_1)}$ and must cancel out ξ iff $s_{N+1} = s_1$

$$\begin{pmatrix} z_{N+1}^+ \\ z_{N+1}^- \end{pmatrix} \mapsto \begin{pmatrix} \xi^{-1} / \omega^+ \\ 1 / \omega^- \end{pmatrix}$$

Then by contraction we pick off the coefficient of the ξ^p term of the series expansion in of $Z_N(\xi, \boldsymbol{z})$ in ξ, denoted here $\mathrm{Coef}[Z_N(\xi, \boldsymbol{z}), \xi, p]$, for $p = 0$:

$$Z_N(\boldsymbol{z}) = \mathrm{Coef}\left[(\xi \quad 1) \cdot \begin{pmatrix} \omega^+ & 0 \\ 0 & \omega^- \end{pmatrix} \cdot \left[\begin{pmatrix} \omega^{++} & \omega^{+-} \\ \omega^{-+} & \omega^{--} \end{pmatrix} \cdot \begin{pmatrix} \omega^+ & 0 \\ 0 & \omega^- \end{pmatrix} \right]^{n-1} \cdot \begin{pmatrix} 1/\omega^+ & 0 \\ 0 & 1/\omega^- \end{pmatrix} \cdot \begin{pmatrix} \xi^{-1} \\ 1 \end{pmatrix}, \xi, 0 \right]$$

$$= \mathrm{Coef}\left[(\xi \quad 1) \cdot \begin{pmatrix} \sqrt{\omega^+} & 0 \\ 0 & \sqrt{\omega^-} \end{pmatrix} \cdot \Omega^N \cdot \begin{pmatrix} \sqrt{\omega^+} & 0 \\ 0 & \sqrt{\omega^-} \end{pmatrix}^{-1} \cdot \begin{pmatrix} \xi^{-1} \\ 1 \end{pmatrix}, \xi, 0 \right]$$

$$= \mathrm{Trace}\left[\Omega^N\right]$$

as claimed.

5.6 Contraction operation in general

A general expression for an elementary *contraction* that identifies two variables s_i and s_j with fugacities z_i and z_j is

$$Z(\boldsymbol{z}) \,|_{z_j \mapsto z_i} = \mathrm{Coef}\left[Z_{\mathrm{full}}\left(\boldsymbol{z}' \mid z'_i \mapsto \xi_i z_i, z'_j \mapsto \xi_i^{-1} \tilde{\omega}_i^{-1}, \{z'_k \mapsto z_k\}\right), \xi, 0 \right]$$

$$= \frac{1}{2\pi i} \int_C \frac{d\xi}{\xi} \, Z_{\mathrm{full}}\left(\boldsymbol{z}' \mid z'_i \mapsto \xi_i z_i, z'_j \mapsto \xi_i^{-1} \tilde{\omega}_i^{-1}, \{z'_k \mapsto z_k\}\right)$$

where $\tilde{\omega}_i$ represents all relative probability factors that are fully redundant between z_i and z_j and C is a contour around the origin within a region where the integrand is analytic. The "Coef" operation finds the coeffecient of ξ^0 in the series expansion of its input function in ξ. Since this operation is linear in its input function, it is a linear functional. Analytically it is the zeroth member of the Inverse Z-Transform sequence of a function, usually denoted "$\mathcal{Z}^{-1}$" rather than "Coef". Contractions on multiple pairs of variables can be defined similarly, without regard to their order.

References

[1] Mjolsness, E., & Yosiphon, G. (2007, January). *Stochastic Process Semantics for Dynamical Grammars*. Annals of Mathematics and Artificial Intelligence, **47**(3-4).

[2] Athreyea, K. B., & Ney, P. E. (1972). *Branching Processes*. Springer-Verlag; Dover.

[3] Dyson, F. (1969). *Existence of a Phase-Transition in a One-Dimensional Ising Ferromagnet*. Comm. Math. Phys., **12**, 91–107.

[4] Mjolsness, E. (2007). *On Cooperative Quasi-Equilibrium Models of Transcriptional Regulation*. Journal of Bioinformatics and Computational Biology, in press.

[5] Monod, J., Wyman, J., & Changeaux, J. P. (1965). *On the nature of allosteric transitions: A Plausible Model*. Journal of Molecular Biology, **12**, 88–118.

[6] Jaeger, J., et al. (2004). *Dynamic control of positional information in the early Drosophila blastoderm*. Nature, **430**, 368–371.

[7] Najdi, T. S., et al. (2006). *Application of a Generalized MWC Model for the Mathematical Simulation of Metabolic Pathways Regulated by Allosteric Enzymes*. Journal of Bioinformatics and Computational Biology, **4**, 335–355.

[8] Shapiro, B. E., et al. (2003). *Cellerator: extending a computer algebra system to include biochemical arrows for signal transduction simulations* . Bioinformatics , **19**, 677–678.