Energy-based Modeling in BioNetGen

John A.P. Sekar,[1,2] Justin S. Hogg,[1] and James R. Faeder[1]

[1]Department of Computational & Systems Biology,
University of Pittsburgh School of Medicine
Pittsburgh PA 15260 USA

[2]Department of Genetics and Genomic Sciences
Mt. Sinai School of Medicine
New York NY 10029 USA

Abstract—**Biochemical processes typically operate on molecular sites (domains, motifs, residues). In rule-based modeling languages such as BioNetGen (BNG), reactions driven by the same set of sites with the same rate law can be represented using a single reaction rule. This leads to compact model representations when sites behave independently, but not when a large number of sites interact cooperatively. Additionally, loops of reactions are constrained by detailed balance equations involving their rate constants, but these constraints have to be enforced manually when specifying rate laws for rules. Here, we introduce the energy-based BioNetGen specification (eBNG), in which models always satisfy detailed balance and cooperative interactions are compactly specified as free energy contributions. We demonstrate this approach using well-known models of molecular cooperativity, such as ligand-induced receptor dimerization and allosteric modulation.**

Keywords—rule-based modeling; reaction network; detailed balance; cooperativity; regulatory complexity.

I. Introduction

The reaction network is a typical mathematical representation of chemical kinetics. The rule-based model is an abstract layer above the reaction network in which chemical species can be grouped based on their molecular structures (“patterns”) and reactions can be grouped into classes (“reaction rules”), such that all reactions in a class have the same rate law formula and are driven by the same patterns [1]. When causal relationships between sites are sparse, a small number of reaction rules can compactly specify a large reaction network [1]. However, rule-based models are limited by the problems of detailed balance and regulatory complexity. By the principle of detailed balance, the net equilibrium flux through a loop of reversible processes must be zero, which constrains rate parameters in the form of a detailed balance equation. These constraints are only enforced manually when specifying rate laws for rules [2]. Regulatory complexity arises due to cooperative interactions between sites. Each combination of sites will regulate a process in a unique manner, leading to as many rules for that process as there are unique site combinations. This inflates the size of the model when assumptions about cooperativity are included [3], [4].

BioNetGen (BNG) is a rule-based framework that has been used to specify many large models [5], [6]. BNG uses a graph syntax for specifying patterns of molecular structures and a graph rewriting approach for specifying reaction rules [6]–[8]. Here, we introduce the energy-based extension to BioNetGen (eBNG) in which kinetic parameters and cooperative effects are expressed compactly as free energy values associated with molecular patterns [3]. In this framework, reaction rate laws are automatically determined from reaction free energies such that detailed balance constraints are always satisfied [3].

II. Background

First we define the basic elements of reaction network models and rule-based models. Then, we discuss the problems of detailed balance and regulatory complexity.

A. Reaction Network Specification

The **chemical species** is a unique chemical entity typically represented by a label such as A, B, or C. The **reaction** is a process that consumes certain species and produces others, such as A+ B→C. The **rate law** is a formula for calculating rate of a reaction, e.g., for the reaction A+ B→C, the rate law can be described using the ordinary differential equation

$$-d[\mathrm{A}]/dt = -d[\mathrm{B}]/dt = d[\mathrm{C}]/dt = k{*}[\mathrm{A}]{*}[\mathrm{B}]$$

where [A], [B], [C] denote concentrations of species A, B, C respectively and *t* denotes time. The **reaction network model** is a set of species labels and a set of reactions on those species, with each reaction having a specified rate law.

B. Rule-based Specification

Briefly, in BioNetGen [5], [6], molecules of different types are linked by bonds to form graphs that represent chemical species. Groups of species can be identified by matching subgraphs called patterns and transformed using graph-rewriting statements called reaction rules. Kinetics is specified by assigning a rate law to each reaction rule.

A **molecule type** has a label and a specified number of components. Each component is of a specific **component type**. Each component type has one or more **states** available to it. If no state is assigned to a component type, a default null state is assumed:

Syntax: MoleculeType(ComponentType ~State ~State, ...)

- B(a) is a molecule type B with a single component of type ‘a’ that takes a default null state.
- A(b~x~y, c) is a molecule type A with two components, one each of component types ‘b’ and ‘c’. {x,y} are states available to ‘b’, whereas ‘c’ has a default null state.

- X(a,a,a) is a molecule type X with three components of type 'a' that has a default null state.

The **fully configured molecule** (or simply, **molecule**) is created from the molecule type by fixing a single state for each component.

- B(a) is a molecule created from molecule type B(a).
- A(b~x,c) and A(b~y,c) are molecules created from molecule type A(b~x~y,c).

The **bond** is a link between a pair of components that can be used to bring two different molecules into proximity. The **species graph** consists of one or more molecules linked by bonds:

Syntax: M_1 **.** M_2 **.** ...,

where M_i = Molecule(Component [~State] [! BondLabel], ...), where [ ] indicates that specification of state or bond label is optional. Two components linked by a bond have the same bond label, whereas unbound components do not have a bond label.

- A(b~x,c) is a species graph with one molecule of type A.
- A(b~x!1,c).B(a!1) is a species graph with one molecule each of types A and B. The bond labeled !1 links components 'b' and 'a'.
- X(a,a,a!1).X(a!1,a,a!2).X(a!2,a,a) is a species graph with three molecules of type X linked by two bonds labeled !1 and !2.

Species graphs enable a rich interpretation of the chemical reaction as a graph rewriting, e.g., the reaction

$$\mathrm{A(b{\sim}x,c) + B(a) \rightarrow A(b{\sim}x!1,c).B(a!1)}$$

represents a graph rewriting operation that adds a bond between components 'b' and 'c'. Henceforth, the terms species and reaction refer to species graphs and their graph rewritings respectively.

The **pattern** is a generalization of the species graph that allows structural features to be omitted or left unspecified. Patterns enable the systematic creation of subgraphs from species graphs:

- A(b~x,c) is a species graph. Omit state 'x' and component 'c' to get pattern A(b).
- A(b~x!1,c).B(a!1) is a species graph. Omit component 'c' and state 'x' to get pattern A(b!1).B(a!1).
- !+ is a special bond label that indicates that one half of the bond is unspecified, e.g., B(a!+).
- !? is a special bond label that indicates that the presence or absence of a bond is left unspecified, e.g., B(a!?).

The pattern p **embeds** into another pattern s, denoted $p < s$, by matching a subgraph of s. The rules underlying pattern isomorphism have been described in the Supplement of [8]. s could be the same pattern, a different pattern or a species graph, e.g., for pattern A(b),

- A(b) < A(b), i.e., into itself,
- A(b) < A(b,c), where A(b,c) is a pattern,
- A(b) < A(b~x,c), where A(b~x,c) is a species.

The species graphs matched by a pattern constitute a **species class**, i.e., A(b) < {A(b~x,c), A(b~y,c)}. The pattern as a match condition is sufficient to define this class, i.e., {species | A(b) < species}. This makes it possible to define a class of species without explicitly specifying its members.

The **match count**, $n(p,s)$, is the number of different ways by which p can embed into s, which can exceed one. For example, X(a) < X(a,a,a) has a match count of three, because the component 'a' on the left may match to any 'a' component on the right. A **symmetric** pattern p is one for which $n(p,p)>1$.

The **reaction rule** is a graph rewriting on patterns, written with the following syntax:

$$\mathrm{RuleLabel:\ RP_1 + RP_2 + ... \rightarrow PP_1 + PP_2 + ...\quad RateConst}$$

where RP_i denotes a reactant pattern and PP_i denotes a product pattern. For reversible processes, ↔ is used and two rate constants are provided. Consider the reaction rule:

$$\mathrm{R:\quad A(b) + B(a) \leftrightarrow A(b!1).B(a!1)} \qquad k_f, k_r$$

Here, the rule has label R, reactant patterns A(b) and B(a), and product pattern A(b!1).B(a!1). It implements the rewriting operation "AddBond" on components 'b' and 'a' in the forward direction with rate constant k_f and the "DeleteBond" operation in the reverse direction with rate constant k_r. A rule may specify more than one rewriting operation of any type. Although eBNG rules support the different types of rewriting operations available in BNG [6], [9], we restrict our discussion here to the rewriting operations of adding/removing bonds and changing component states.

The reactant patterns define species classes and the rule defines a graph rewriting on the reactant patterns. Transitively, the reaction rule defines equivalent rewritings on all matched species graphs, e.g., since A(b) < {A(b~x,c), A(b~y,c)} and B(a) < {B(a)}, R matches the reactions

$$\mathrm{Rxn_1:\quad A(b{\sim}x,c) + B(a) \rightarrow A(b{\sim}x!1,c).B(a!1)} \qquad k$$
$$\mathrm{Rxn_2:\quad A(b{\sim}y,c) + B(a) \rightarrow A(b{\sim}y!1,c).B(a!1)} \qquad k$$

The reactions matched by a rule constitute an equivalence class, i.e., R < {Rxn_1, Rxn_2}. The reaction rule as a match condition is sufficient to define this class, i.e., {reaction | R < reaction}. Propagating the rate function assigned to the rule to all matching reactions defines a **reaction class.** This makes it possible to specify and parameterize reactions as a class without building each reaction explicitly. The **rule-based model** is a set of seed species and a set of reaction rules with a rate law specified for each reaction rule.

Network generation is the process of expanding a set of reaction rules to build all matching reactions. This is performed by iteratively applying reaction rules to a species set to generate new species and reactions and updating the initial sets. The output is a reaction network with kinetics that is exactly equivalent to the rule-based model, with symmetry and multiplicity factors handled automatically [3], [7]. Network generation allows the automatic construction of large reaction networks from small sets of reaction rules [6].

C. The problem of Detailed Balance

The problem of detailed balance arises when reversible reactions (or reaction rules) are specified as if they are independently determined rather than mutually constrained by thermodynamic cycles. For example, consider a cycle of reversible reactions, such as A↔B↔C↔A, where the forward and reverse rate constants at each branch of the loop are (k_{f1},k_{r1}), (k_{f2},k_{r2}), and (k_{f3},k_{r3}) respectively. The equilibrium constants at each branch of the loop are $K_1 = k_{f1}/k_{r1}$, $K_2 = k_{f2}/k_{r2}$, and $K_3 = k_{f3}/k_{r3}$ respectively. By the principle of detailed balance, the net flux around a cycle must be zero at equilibrium, which constrains K_1, K_2 and K_3 as

$$K_1 * K_2 * K_3 = 1$$

Typically, these constraints are manually imposed when building a rule-based model [2] and can be violated by a careless modeler. Generating the reaction network from the rule-based model and verifying detailed balance by cycle detection is computationally hard for very large networks [10]. This motivates the need for a specification that naturally constrains detailed balance without user direction [1].

D. The problem of Regulatory Complexity

The rule-based specification takes advantage of independence relationships between sites to minimize the amount of information needed to build a reaction class. This enables compact models when sites are mostly independent. However, when sites interact cooperatively, each combination of sites that contributes a unique kinetic effect to a process has to be modeled as a separate rule, so very little compression is achieved when casting the rules that represent the system [3], [4]. We demonstrate this using a simple two-component, two-state system with the following reaction network:

$$\begin{array}{ll} \mathrm{Rxn}_1\text{: A(b}\sim\text{0,c}\sim\text{0)} \leftrightarrow \text{A(}\underline{\text{b}\sim\text{1}}\text{,c}\sim\text{0)} & k_{f1}, k_{r1} \\ \mathrm{Rxn}_2\text{: A(b}\sim\text{0,c}\sim\text{1)} \leftrightarrow \text{A(}\underline{\text{b}\sim\text{1}}\text{,c}\sim\text{1)} & k_{f2}, k_{r2} \end{array}$$

$$\begin{array}{ll} \mathrm{Rxn}_3\text{: A(b}\sim\text{0,c}\sim\text{0)} \leftrightarrow \text{A(b}\sim\text{0,}\underline{\text{c}\sim\text{1}}\text{)} & k_{f3}, k_{r3} \\ \mathrm{Rxn}_4\text{: A(b}\sim\text{1,c}\sim\text{0)} \leftrightarrow \text{A(b}\sim\text{1,}\underline{\text{c}\sim\text{1}}\text{)} & k_{f4}, k_{r4} \end{array}$$

(underline emphasizes the rewriting operation). The detailed balance constraint for this system is the equation:

$$(k_{f1}/k_{r1})*(k_{f4}/k_{r4}) = (k_{f3}/k_{r3})*(k_{f2}/k_{r2})$$

Consider building a rule-based model of this system assuming either independence or cooperativity between sites 'b' and 'c'. Under the cooperativity assumption, the rate at which 'b' is activated or deactivated depends on the state of 'c' and *vice versa*. Modeling the system requires building a separate rule for each of the four reactions. On the other hand, under the independence assumption, the rate at which 'b' is activated or deactivated is independent of the state of 'c', so the rate constants of reactions Rxn_1 and Rxn_2 would be set to identical values. Similarly, rate constants for reactions Rxn_3 and Rxn_4 would have identical values also. This can be concisely expressed using 2 rules and 4 independent rate constants:

$$\begin{array}{ll} \mathrm{R_b}\text{: A(b}\sim\text{0)} \leftrightarrow \text{A(}\underline{\text{b}\sim\text{1}}\text{)} & k_{fb}, k_{rb} \\ \mathrm{R_c}\text{: A(c}\sim\text{0)} \leftrightarrow \text{A(}\underline{\text{c}\sim\text{1}}\text{)} & k_{fc}, k_{rc} \end{array}$$

Starting with the independence assumption and then including cooperativity assumptions requires a significant rewrite and expansion of the model architecture: from 2 rules and 4 independent parameters to 4 rules, 8 parameters and 1 detailed balance constraint. The discrepancy grows combinatorially with the number of processes and states involved, e.g., for a three-component two-state system, the independence assumption requires 3 rules and 6 independent parameters, whereas the cooperative assumption requires 12 rules, 24 parameters and 5 constraints. This motivates the need for a specification in which cooperative assumptions can be expressed compactly [3].

III. Methods

A. Principle

The goal is to enable a rule-based specification in which network generation always leads to networks that satisfy detailed balance. First, we translate the problem defined in Section II-C to the free energy space. The free energy change associated with a reaction ΔG^0 is related to the equilibrium rate constant K by the equation

$$\Delta G^0 = -R*T*\ln(K)$$

where R is the universal gas constant and T is the temperature. For convenience, we will consider all energies to have been scaled by $R*T$, so,

$$\Delta G = \Delta G^0 / (R*T) = -\ln(K)$$

For the reaction loop A↔B↔C↔A from Section II-C above, the detailed balance constraint can be restated as

$$\Delta G_1 + \Delta G_2 + \Delta G_3 = 0$$

where $\Delta G_1 = -\ln(K_1)$, $\Delta G_2 = -\ln(K_2)$, and $\Delta G_3 = -\ln(K_3)$ are the free energy changes associated with each reaction in the loop. Setting ΔG_1, ΔG_2 and ΔG_3 independently for each reaction would break detailed balance. On the other hand, if they were determined as follows:

$$\begin{aligned} \Delta G_1 &= \Delta G_\mathrm{B} - \Delta G_\mathrm{A} \\ \Delta G_2 &= \Delta G_\mathrm{C} - \Delta G_\mathrm{B} \\ \Delta G_3 &= \Delta G_\mathrm{A} - \Delta G_\mathrm{C} \end{aligned}$$

where ΔG_A, ΔG_B and ΔG_C are the free energies of formation of species A, B and C respectively, then the detailed balance constraint is satisfied for all values of ΔG_A, ΔG_B and ΔG_C.

Thus, one solution to the detailed balance problem is to specify kinetics indirectly using free energies of formation of species as independent parameters. However, in a rule-based framework, species and reactions are not known *a priori,* and the reaction network is generated *after* the kinetic parameters are specified. The challenge then is to

1. specify free energies of formation of species without populating the species set;
2. determine rate constants compatible with the computed reaction free energies; and
3. embed model kinetics using energy values.

B. Specifying Free Energies of Formation

An **energy pattern** is a pattern p assigned an **energy value**, denoted $E[p]$. Let Ω be the set of energy patterns defined for a model. For p not in Ω, $E[p] = 0$. The **contribution** of an energy pattern p to a species graph s, denoted $E[p,s]$, is computed as

$$E[p,s] = (n(p,s) / n(p,p))*E[p]$$

The **free energy of formation of a species** s, denoted $E[s]$, is the sum of contributions from energy patterns, i.e.,

$$E[s] = \Sigma_{p\in\Omega} E[p,s]$$

Consider the example in Section II-D that uses molecule type A(b~0~1,c~0~1). Consider the energy pattern set

$$p_1\text{=A(b\~{}1)}, p_2\text{=A(c\~{}1)}, p_3\text{=A(b\~{}1,c\~{}1)}$$

The energies for the species containing A are

$$E[\text{A(b\~{}0,c\~{}0)}] = 0$$

$$E[\text{A(b\~{}1,c\~{}0)}] = E[p_1]$$

$$E[\text{A(b\~{}0,c\~{}1)}] = E[p_2]$$

$$E[\text{A(b\~{}1,c\~{}1)}] = E[p_1] + E[p_2] + E[p_3]$$

The **free energy of a reaction** μ, denoted ΔG_μ, is the sum of the free energies of the product species minus the sum of the free energies of the reactant species, i.e.,

$$\Delta G_\mu = \Sigma_{s\in Pr(\mu)} E[s] - \Sigma_{s\in Re(\mu)} E[s]$$

where $Re(\mu)$ and $Pr(\mu)$ denote the reactant and product species for the reaction μ. For the reactions

$$\text{Rxn}_1\text{: A(b\~{}0,c\~{}0)} \leftrightarrow \text{A(b\~{}1,c\~{}0)}$$
$$\text{Rxn}_2\text{: A(b\~{}0,c\~{}1)} \leftrightarrow \text{A(b\~{}1,c\~{}1)},$$

the respective reaction free energies are given by

$$\Delta G_1 = E[\text{A(b\~{}1,c\~{}0)}] - E[\text{A(b\~{}0,c\~{}0)}]$$

$$\Delta G_2 = E[\text{A(b\~{}1,c\~{}1)}] - E[\text{A(b\~{}1,c\~{}0)}]$$

Substituting energies calculated from energy pattern matches gives

$$\Delta G_1 = (E[p_1]) - (0) = E[p_1]$$

$$\Delta G_2 = (E[p_1] + E[p_2] + E[p_3]) - (E[p_2]) = E[p_1] + E[p_3]$$

Both reactions effect a change in the state of 'b', but under two different contexts that depend on the state of 'c'. The second reaction creates a co-occurrence of b~1 and c~1, which leads to an additional contribution $E[p_3]$ from the energy pattern p_3. $E[p_3]$=0 represents the independence assumption that results in $\Delta G_1=\Delta G_2$, whereas $\mathrm{E}[p_3]\neq 0$ represents the cooperative assumption that results in $\Delta G_1\neq\Delta G_2$.

C. Types of Energy Patterns

Energy patterns can be broadly classified as **Type I** patterns, which specify exactly one fundamental structure, and **Type II** patterns, which specify combinations of structures. Examples of Type I patterns are A(b), which specifies an unbound component, A(b!1).B(a!1), which specifies a bond, and A(b~x!?), which specifies a component state. Note that without the '!?' wildcard, the last pattern would specify both the state and the binding status of the component (unbound), and so would be a Type II pattern. Examples of Type II patterns are A(b~x,c), which specifies that component 'b' is unbound and has state 'x' and component 'c' is unbound, and A(b~x!1,c).B(a!1), which specifies that component 'c' is unbound, components 'b' and 'a' are bound to each other, and component 'b' has state 'x'.

The energy value of a Type I pattern is effectively the free energy of formation in the absence of influences from other structures. Choosing some Type I patterns to be default zero-energy states while assigning non-zero values to others is equivalent to choosing the default equilibrium constants for each process. The energy value of a Type II pattern is a modifier that lowers or increases the free energy of a species containing that pattern (as seen in Section III-B). Setting a negative value stabilizes any species matching the pattern and leads to an increase in the equilibrium constants representing its formation, whereas setting a positive value achieves the opposite. Setting the value to zero results in no contribution. Type II patterns can specify any number and combination of structures so cooperative interactions can be defined to any arbitrary complexity.

D. Determining Rate Constants

The linear transition state theory provides a basis for computing reaction rate constants k_f and k_r constrained by the reaction free energy $\Delta G_{\text{rxn}} = -\ln(k_f/k_r)$ [11]. It assumes an intermediary transition state whose free energy changes are a linear combination of the free energy changes of the reactants and products. Let $\{G_{R,0}, G_{P,0}, G_{TS,0}\}$ and $\{G_{R,1}, G_{P,1}, G_{TS,1}\}$ be the free energies of reactant, product and transition state under independent system variables denoted by subscripts 0 and 1 respectively. If we treat subscript 0 as the reference state, then the free energies of formation for reactant, product and transition state are respectively,

$$\Delta G_R = G_{R,1} - G_{R,0}$$
$$\Delta G_P = G_{P,1} - G_{P,0}$$
$$\Delta G_{TS} = G_{TS,1} - G_{TS,0}$$

The free energy associated with the reaction in each system state is given by

$$\Delta G_0 = G_{P,0} - G_{R,0}$$
$$\Delta G_1 = G_{P,1} - G_{R,1}$$

The change in the reaction free energy with respect to the reference state is given by

$$\Delta G_{\text{rxn}} = \Delta G_1 - \Delta G_0 = \Delta G_P - \Delta G_R$$

From the linear transition state assumption,

$$\Delta G_{TS} = \varphi*\Delta G_P + (1-\varphi)*\Delta G_R$$

The activation energies in the forward direction is the energy difference between transition state and reactant. Similarly, the activation energy in the reverse direction is the energy difference between transition state and product. So,

$$E_{Af,0} = G_{TS,0} - G_{R,0}$$
$$E_{Af,1} = G_{TS,1} - G_{R,1}$$
$$E_{Ar,0} = G_{TS,0} - G_{P,0}$$
$$E_{Ar,1} = G_{TS,1} - G_{P,1}$$

Substituting from above, the changes in the forward and reverse activation energies can be related to the changes in the reaction free energies as:

$$\Delta E_{Af} = E_{Af,1} - E_{Af,0} = \Delta G_{TS} - \Delta G_R = \varphi * \Delta G_{\mathrm{rxn}}$$

$$\Delta E_{Ar} = E_{Ar,1} - E_{Ar,0} = \Delta G_{TS} - \Delta G_P = (\varphi - 1) * \Delta G_{\mathrm{rxn}}$$

Integrating and introducing constants E_A and φ, we get

$$E_{Af} = E_A + \varphi * \Delta G_{\mathrm{rxn}}$$

$$E_{Ar} = E_A + (\varphi - 1) * \Delta G_{\mathrm{rxn}}$$

From the Arrhenius theory of reaction rates,

$$k_f = \sigma_f * \exp(- E_{Af})$$

$$k_r = \sigma_r * \exp(- E_{Ar})$$

where σ_f and σ_r are scaling constants with appropriate units that relate the dimensionless energy terms to kinetic rate constants. For convenience, we can absorb their actual numeric values into the activation energy terms and set the scaling constants to 1. Then,

$$k_f = \exp(- (E_A + \varphi * \Delta G_{\mathrm{rxn}}))$$

$$k_r = \exp(-(E_A + (\varphi - 1) * \Delta G_{\mathrm{rxn}})$$

Thus it is possible to calculate k_f and k_r given ΔG_{rxn} and independent parameters E_A and φ. E_A scales the rate constant values relative to the energy values. φ takes a value between 0 and 1 and controls how differences in ΔG_{rxn} contribute to differences in k_f and k_r.

E. Energy Rules

In the original BNG formalism [8], a reaction class was defined as one in which members of the class shared the same rate law function. In the eBNG approach, the rate laws are determined automatically from ΔG_{rxn}, E_A and φ (Section III-D), with ΔG_{rxn} computed automatically from counting matches to energy patterns (Section III-B), so it suffices to specify E_A and φ as attributes of a reaction class. An energy rule has the following syntax:

RuleLabel: $RP_1 + RP_2 + ... \leftrightarrow PP_1 + PP_2 + ...$ Arrhenius(φ,E_A)

The Arrhenius keyword indicates to BioNetGen that, during network generation, the calculated reaction free energy must be used in determining the rate constants for any reaction derived from this rule. φ and E_A may be specified either as numeric expressions or local functions evaluated over individual reactions. Typically, it is sufficient to build energy-based rules representing elementary graph rewriting operations, e.g.,

R1: A(b) + B(a) ↔A(b!1).B(a!1)   Arrhenius(φ_1,E_{A1})

R2: A(b~0) ↔A(b~1)   Arrhenius(φ_2,E_{A2})

which carry out binding/unbinding and state change respectively.

F. Choosing E_A, φ and energy values

Suppose kinetic parameters (k_f, k_r or K) are known for a few instances of reactions. It is possible to use this information to parameterize the eBNG model:

1. Choose a subset of Type I patterns to be default states with zero energies.
2. If for a Type I pattern *p,* a default *K* is known or can be calculated (e.g., $K = k_f/k_r$), set $E[p] = -\ln(K)$.
3. Setting φ=0 assumes that the forward rate constant is unchanged for all reactions within a class and free energy contributions only modify the reverse rate constant. Setting φ=1 assumes the converse. Standard practice varies, but in the absence of other information it may be safest to set φ=0.5.
4. If both k_f and k_r are known for a particular reaction instance, or can be calculated (e.g., $k_f = k_r$*K), and φ is known or assumed, set

   $$E_A = -[\varphi * \ln(k_r) + (1 - \varphi) * \ln(k_f)]$$

   Alternatively if E_A is known or assumed, set

   $$\varphi = (E_A + \ln(k_f))/(\ln(k_f) - \ln(k_r))$$

5. For each relevant Type II pattern *p,* associate a value α_p and set $E[p] = -\ln(\alpha_p)$.

G. Implementation

The BioNetGen model file includes text blocks for specifying parameters, molecule types, seed species and reaction rules [6]. An additional block called "energy patterns" is used to specify the energy pattern set Ω [3]. The reaction rules block contains both reaction rules of the original form as well as energy-based rules. Network generation and simulation methods are called using "action" commands [6]. During network generation (as in Section II-B-8), for each reaction generated from an energy rule,

1. free energies of reactant and product species are computed from energy patterns as in Section III-B,
2. reaction free energy is computed from species free energies as in Section III-B,
3. forward and reverse rate constants are computed according to Section III-C.

Since the energy of each species is calculated explicitly, the calculated reaction free energies always satisfy detailed balance (as described in Section III-A). Thus, the portion of the network that was derived from energy-based rules will always have rate constants that satisfy detailed balance relationships.

IV. Examples

We explore classical examples of kinetic models with a focus on cooperative interactions and show how they can specified in eBNG.

A. Ligand-induced Dimerization

The epidermal growth factor receptor is a well-known receptor system in which receptor dimerization is induced by ligand binding [12]. Consider a model with 6 species: free ligand L, free monomer R, ligated monomer LR, unligated dimer RR, singly ligated dimer LRR and doubly ligated dimer LRRL. The reaction network below captures all possible reactions in the system:

$$\begin{array}{ccccc} \text{L, R, R, L} & \leftrightarrow_1 & \text{LR, R, L} & \leftrightarrow_1 & \text{LR, LR} \\ \updownarrow_4 & & \updownarrow_5 & & \updownarrow_6 \\ \text{L, RR, L} & \leftrightarrow_2 & \text{LRR, L} & \leftrightarrow_3 & \text{LRRL} \end{array}$$

Building the model using kinetic classes requires each reaction be modeled by a separate rule, so it requires 6 reversible rules (subscripts 1-6 above) and 2 detailed balance constraints. However, building an energy-based model only requires 2 rules, one each for ligand-binding and dimerization. Using molecule types L(r) for ligand and R(L,d) for receptor, the energy-based rules are

R_{lig}: L(r) + R(L) ↔ L(r!1).R(L!1) Arrhenius(φ_{LR},$E_{A\text{-}LR}$)
R_{dim}: R(d) + R(d) ↔ R(d!1).R(d!1) Arrhenius(φ_{RR},$E_{A\text{-}RR}$)

The energy patterns used are

p_{LR}= L(r!1).R(L!1),
p_{RR}= R(d!1).R(d!1),
p_{LRR}= L(r!1).R(L!1,d!2).R(d!2),
p_{LRRL}= L(r!1).R(L!1,d!2).R(d!2,L!3).L(r!3),

p_{LR} and p_{RR} are Type I patterns representing the ligand-receptor bond and the receptor dimer bond respectively. p_{LRR} and p_{LRRL} are Type II patterns. p_{LRR} represents co-occurrence of ligand bond and dimer bond. p_{LRRL} represents co-occurrence of two ligand bonds and the dimer bond. Following the approach laid out in Section III-F, the energy pattern set Ω is defined as

$$E[p_{LR}] = -\ln (K_{\text{LR}}),\ E[p_{RR}] = -\ln (K_{\text{RR}})$$

$$E[p_{LRR}] = -\ln (\alpha),\ E[p_{LRRL}] = -\ln (\beta)$$

Thus, the kinetics of the model have been expressed using 8 independent parameters { φ_{LR}, $E_{\text{A-RR}}$, φ_{RR}, $E_{\text{A-RR}}$, K_{LR}, K_{RR}, α, β }. During network generation, the contribution of each energy pattern to each species is calculated according to Section III-B. Since p_{RR} and p_{LRRL} are symmetric patterns, their match counts are modified by a factor of 0.5. Free ligand L and unligated monomer R have zero matches to energy patterns, so

$$E[\text{L}] = 0,\ E[\text{R}] = 0$$

Ligated monomer LR matches p_{LR} once. Therefore, its energy is computed to be the negative logarithm of K_{LR}.

$$E[\text{LR}] = 1{*}E[p_{LR}] = -\ln (K_{LR})$$

Unligated dimer RR matches p_{RR} twice, but is corrected by the self-embedding term. Its energy is computed to be the negative logarithm of K_{RR}.

$$E[\text{RR}] = 2{*}0.5{*}E[p_{RR}] = -\ln (K_{RR})$$

Singly ligated dimer LRR matches p_{LR} and p_{LRR} once each and p_{RR} twice, with p_{RR} matches corrected by 0.5, so

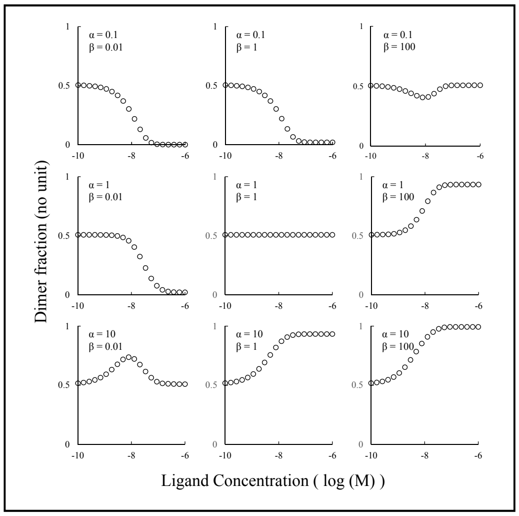
$$E[\text{LRR}] = 1{*}E[p_{LR}]+ 2{*}0.5{*}E[p_{RR}]+ 1{*}E[p_{LRR}]$$
$$= -\ln (\alpha K_{RR}K_{LR})$$

Doubly ligated dimer LRRL matches all four pattern twice each, with p_{RR} and p_{LRRL} matches corrected by 0.5, so

$$E[\text{LRRL}] = 2{*}E[p_{LR}] + 2{*}0.5{*}E[p_{RR}] + 2{*}E[p_{LRR}] + 2{*}0.5{*}E[p_{LRRL}]$$
$$= -\ln (\alpha^2\beta K_{RR}\, K_{LR}^2)$$

Depending on the values set for α and β, the model architecture enables multiple ways in which dimerization can respond to ligand dose (Fig. 1). When α=β=1, the dimer fraction at equilibrium is independent of ligand concentration. For some settings, dimer fraction exhibits an overall decrease or increase with ligand concentration. For other settings, the dose response is more complicated and can have a maximum or minimum value at some intermediate ligand concentration. In the epidermal growth factor system, an increase in dimerization at certain ligand concentration regimes activates downstream pathways associated with cell growth and cancer [13].

Fig. 1. Ligand-induced dimerization. (*x-axis:* log of ligand concentration in molar units, *y-axis:* fraction of receptors in dimers).



B. Allosteric Modulation

The Monod-Wyman-Changeux model of hemoglobin is a classic example of a biochemical kinetic process regulated by cooperative interactions [14], [15]. Hemoglobin is an oxygen transport molecule with 4 globin subunits that can individually bind and release oxygen. Each subunit can be in a relaxed (R) state that binds oxygen with high affinity or a tense state (T) that binds with low affinity. All four subunits are assumed to transition between R and T states in a concerted fashion. The 2,3-bisphosphoglycerate molecule (BPG) binds hemoglobin at a separate binding site and modulates oxygen affinity, a phenomenon called allosteric modulation. In addition to the hemoglobin molecule being in R/T states and BPG-bound vs. unbound states, the hemoglobin molecule can be bound to between 0 and 4 oxygen molecules at a time. This leads to a complex reaction network with 22 species and 36 reversible reactions. To specify the model in eBNG, we use molecule types H(m~R~T,g,g,g,g,b), Ox(h) and B(h) to represent hemoglobin, oxygen and BPG respectively. By using four components named ‘g’, we represent the assumption that all four subunits have identical properties. We only need three energy-based rules:

R_{RT}: H(m~R) ↔ H(m~T) Arrhenius(φ_{RT},$E_{A\text{-}RT}$)
R_{Ox}:H(g) + Ox(h) ↔ H(g!1).Ox(h!1) Arrhenius(φ_{Ox},$E_{A\text{-}Ox}$)
R_B: H(g) + B(h) ↔ H(g!1).B(h!1) Arrhenius(φ_B,$E_{A\text{-}B}$)

The φ and E_A terms affect the rates of individual forward and reverse reactions but not the final equilibrium concentrations of species, so we set them to default values. To define the default equilibrium constants, we use Type I energy patterns:

$$E[\text{ H(m\textasciitilde T) }] = -\ln(K_{RT})$$

$$E[\text{ H(g!1).Ox(h!1) }] = -\ln(K_{Ox})$$

$$E[\text{ H(b!1).B(h!1) }] = -\ln(K_B)$$

This sets K_{RT} as the equilibrium constant for the R to T transition in the absence of oxygen and BPG, K_{Ox} as the expected equilibrium constant for oxygen-binding when sites are assumed to be independent, and K_B as the equilibrium constant for binding BPG in the absence of oxygen. To express cooperativity between the processes, we use Type II patterns:

$$E[\text{ H(m\textasciitilde T,g!1).Ox(h!1) }] = -\ln(\alpha_T)$$

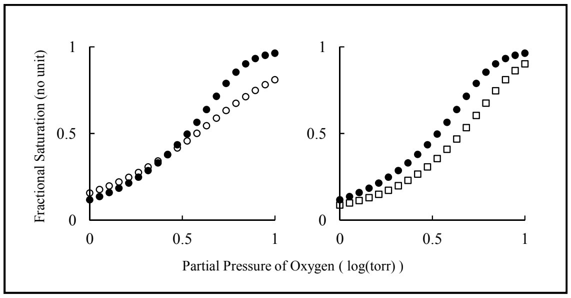
$$E[\text{ H(m\textasciitilde R,g!1).Ox(h!1) }] = -\ln(\alpha_R)$$

$$E[\text{ H(m\textasciitilde T,b!1).B(h!1) }] = -\ln(\beta_T)$$

$$E[\text{ H(m\textasciitilde R,b!1).B(h!1) }] = -\ln(\beta_R)$$

Under this specification, the T and R states have oxygen binding equilibrium constants $\alpha_T K_{Ox}$ and $\alpha_R K_{Ox}$ respectively and BPG binding equilibrium constants $\beta_T K_B$ and $\beta_R K_B$ respectively. Setting α and β terms to different values allows us to explore the functional role of cooperativity in this system (Fig. 2). In the independence assumption, all α and β terms are set to one and the model predicts a graded response for the fractional saturation of oxygen-binding sites with increase in oxygen concentration. Changing the R-state affinity for oxygen to be significantly higher than the T-state affinity predicts the experimentally observed switch-like response that is more suitable for oxygen transport. The allosteric modulator BPG stabilizes the low affinity T-state preferentially over the R-state and this has the experimentally observed effect of shifting the binding curve to the right when BPG concentration is increased.

Fig. 2. Allostery in Hemoglobin. (*x-axis:* log of partial pressure of oxygen, *y-axis:* fractional saturation of globin subunits, ○-independence assumption $\alpha_T = \alpha_R = \beta_T = \beta_R = 1$, ●-cooperative oxygen binding α_T=0.2, α_R=8.4, $\beta_T = \beta_R = 1$, □-modulation by BPG, α_T=0.2, α_R=8.4, β_T = 1, β_R = 0.022)



C. Hierarchical Model Construction

Assigning energy values to increasingly large patterns corresponds directly to making increasingly complex assumptions about the nature of cooperativity between sites [3]. This allows model assumptions to be structured hierarchically, which in turn allows fitting a series of models to experimental data and selecting the one with the smallest number of assumptions that also recapitulates the data. Here, we show an example of a hierarchical model construction in eBNG. Consider a model with three processes:

R1: A(b) + B(a) ↔ A(b!1).B(a!1) Arrhenius(φ_1,E_{A1})
R2: A(b~x) ↔ A(b~y) Arrhenius(φ_2,E_{A2})
R3: A(c) + C(a) ↔ A(c!2).C(a!2) Arrhenius(φ_3,E_{A3})

We construct energy patterns in three sets. Set 1 consists of

A(b~y!?), A(b!1).B(a!1), A(c!2).C(a!2)

These patterns represent the individual structures of ‘y’-state, A-B bond and A-C bond respectively. They are assigned non-zero energy values.

Set 2 consists of

A(b~y!1).B(a!1)
A(b~y!?,c!2).C(a!2)
A(b!1,c!2).B(a!1).C(a!2).

These patterns represent structural combinations of size 2: (y-state, A-B), (y-state, A-C) and (A-B, A-C) respectively. They are assigned energy values of the form b_1*x, where x is some non-zero value and b_1 is a Boolean parameter that can take values 0 or 1.

Set 3 consists of

A(b~y!1,c!2).B(a!1).C(a!2)

This pattern represents a structural combination of size 3: (y-state, A-B, A-C). It is assigned an energy value of the form b_2*x respectively, where x is non-zero b_2 is a Boolean parameter.

Different settings of b_1 and b_2 correspond to models with different levels of assumptions about cooperativity. When b_1=0 and b_2=0, there are no cooperative interactions in the model. When b_1=1 and b_2=0, the model has pairwise cooperative interactions between the processes. When b_1=1 and b_2=1, the model has pairwise cooperative interactions as well as a third-order cooperative interaction involving all three processes. As mentioned previously, this allows building a series of models with different sets of assumptions that can be fit to the same experimental data. Using techniques like LASSO regression [16] and statistical measures like the Bayes factor [17], it is possible to select for the most parsimonious model, i.e., the one with the fewest assumptions about cooperativity that can also satisfactorily explain the data.

D. Phosphocatalysis

In eBNG, it is possible to model phosphorylation by explicitly taking into account the energy balance of the cell [18]. For example, consider the energy-based rules

R1: S(e) + Enz(s) ↔ S(e!1).Enz(s!1)   Arrhenius(φ_1,E_{A1})

R2: S(e~0!1).Enz(s!1) + ATP ↔ S(e~P!1).Enz(s!1) + ADP
Arrhenius(φ_2,E_{A2})

The energy pattern set is devised as follows: *E*[S(e~P)] and *E*[ATP] can be set to positive values to denote high-energy states. *E*[S(e!1).Enz(s!1)] can be set to a negative value to stabilize the enzyme-substrate interaction. *E*[S(e~P!1).Enz(s!1)], when set to a positive value destabilizes the enzyme-product complex after catalysis, and when set to a negative value models inhibition of enzyme by product. Also, ATP and ADP species concentrations can be set to constant unchanging values to model a static ATP/ADP ratio, but allowing them to vary can be used to model altered metabolic states such as starvation [18].

V. Discussion

In summary, we have presented a concise formalism and implementation for describing reaction kinetics in terms of free energies in the BioNetGen language. Our method builds on work by Ollivier *et al.* [4], who developed a rule-based network generation approach for allosteric transitions based on pairwise interactions between components within molecules, and Danos *et al.* [19], who generalized free-energy accounting to a pattern matching framework. We have developed a formal pattern-based system within the overall framework of the BioNetGen language, and, going beyond the previous work of Danos *et al.* [19], we provide an implementation called eBNG that is incorporated into BioNetGen version 2.2 [5]. This rule-based free-energy accounting system enables automated network generation and handling of detailed balance constraints, and we have shown that it can be used to compactly specify systems with cooperative interactions up to high order.

Acknowledgments

We gratefully acknowledge support for this work from the National Institutes of Health under award Nos. P41GM103712, R01AI107825, R01GM115805, and P01HL114453. We thank Sanjana Gupta and Kaiser Loell for helpful comments on the manuscript.

References

[1] L. A. Chylek, L. A. Harris, C.-S. Tung, J. R. Faeder, C. F. Lopez, and W. S. Hlavacek, “Rule-based modeling: a computational approach for studying biomolecular site dynamics in cell signaling systems.,” *Wiley Interdiscip. Rev. Syst. Biol. Med.*, vol. 6, no. 1, pp. 13–36, Sep. 2013.

[2] J. A. P. Sekar and J. R. Faeder, “Rule-Based Modeling of Signal Transduction: A Primer,” *Methods Mol. Biol.*, vol. 880, pp. 139–218, Jan. 2012.

[3] J. S. Hogg, “Advances in Rule-Based Modeling: Compartments, Energy, and Hybrid Simulation, with Application to Sepsis and Cell Signaling,” University of Pittsburgh School of Medicine, 2013.

[4] J. F. Ollivier, V. Shahrezaei, and P. S. Swain, “Scalable rule-based modelling of allosteric proteins and biochemical networks,” *PLoS Comput. Biol.*, vol. 6, 2010.

[5] L. A. Harris, J. S. Hogg, J.-J. Tapia, J. A. P. Sekar, S. Gupta, I. Korsunsky, A. Arora, D. Barua, R. P. Sheehan, and J. R. Faeder, “BioNetGen 2.2: advances in rule-based modeling,” *Bioinformatics*, vol. 32, no. 21, pp. 3366–3368, Nov. 2016.

[6] J. R. Faeder, M. L. Blinov, and W. S. Hlavacek, “Rule-based modeling of biochemical systems with BioNetGen.,” *Methods Mol. Biol.*, vol. 500, pp. 113–67, Jan. 2009.

[7] M. L. Blinov, J. Yang, J. R. Faeder, and W. S. Hlavacek, “Graph theory for rule-based modeling of biochemical networks,” in *Transactions on Computational Systems Biology VII*, vol. 4230, 2006, pp. 89–106.

[8] J. S. Hogg, L. A. Harris, L. J. Stover, N. S. Nair, and J. R. Faeder, “Exact hybrid particle/population simulation of rule-based models of biochemical systems.,” *PLoS Comput. Biol.*, vol. 10, no. 4, p. e1003544, Apr. 2014.

[9] L. A. Harris, J. S. Hogg, and J. R. Faeder, “Compartmental rule-based modeling of biochemical systems,” in *Proceedings of the 2009 Winter Simulation Conference (WSC)*, 2009, pp. 908–919.

[10] J. Yang, W. J. Bruno, W. S. Hlavacek, and J. E. Pearson, “On imposing detailed balance in complex reaction mechanisms.,” *Biophys. J.*, vol. 91, no. 3, pp. 1136–41, Aug. 2006.

[11] J. E. Leffler, “Parameters for the Description of Transition States.,” *Science*, vol. 117, no. 3039, pp. 340–1, Mar. 1953.

[12] C. Wofsy, B. Goldstein, K. Lund, and H. Wiley, “Implications of epidermal growth factor (EGF) induced egf receptor aggregation,” *Biophys. J.*, vol. 63, no. 1, pp. 98–110, Jul. 1992.

[13] A. Citri and Y. Yarden, “EGF-ERBB signalling: towards the systems level.,” *Nat. Rev. Mol. Cell Biol.*, vol. 7, no. 7, pp. 505–16, Jul. 2006.

[14] J. Monod, J. Wyman, and J. Changeux, “On the nature of allosteric transitions: a plausible model.,” *J. Mol. Biol.*, vol. 12, pp. 88–118, May 1965.

[15] T. Yonetani, S. Park, A. Tsuneshige, K. Imai, and K. Kanaori, “Global allostery model of hemoglobin. Modulation of O2 affinity, cooperativity, and Bohr effect by heterotropic allosteric effectors,” *J. Biol. Chem.*, vol. 277, no. 37, pp. 34508–34520, 2002.

[16] R. Tibshirani, “Regression shrinkage and selection via the lasso: a retrospective,” *J. R. Stat. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, Jun. 1996.

[17] R. E. Kass and A. E. Raftery, “Bayes factors,” *J. Am. Stat. Assoc.*, vol. 90, no. 430, pp. 773–795, 1995.

[18] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell, 4th Edition*. Garland Science, 2002.

[19] V. Danos, R. Harmer, and R. Honorato-Zimmer, “Thermodynamic graph-rewriting,” in *Lect. Notes Comp. Sci.*, 2013, vol. 8052, pp. 380–394.