Domain-Oriented Reduction of Rule-Based Network Models

Nikolay M. Borisov\textsuperscript{1,2}, Alexander S. Chistopolsky\textsuperscript{2}, James R. Faeder\textsuperscript{3*} and Boris N. Kholodenko\textsuperscript{1*}

\textsuperscript{1}Department of Pathology, Anatomy and Cell Biology, Thomas Jefferson University, 1020 Locust St., Philadelphia, PA 19107, USA
\textsuperscript{2}State Research Center – Institute of Biophysics, 46, Zhivopisnaya St., Moscow, 123182, Russia
\textsuperscript{3}Department of Computational Biology, University of Pittsburgh School of Medicine, Pittsburgh, PA15260, USA.

Running head: Reduction of Rule-Based Models
E-mail addresses of the authors: Nikolay Borisov nikolay.borisov@jefferson.edu
Alexander Chistopolsky eroxi@savelovo.net
James Faeder faeder@pitt.edu
Boris Kholodenko boris.kholodenko@jefferson.edu

*Correspondence should be addressed to:
James R. Faeder
Department of Computational Biology
3082 Biomedical Science Tower 3
University of Pittsburgh School of Medicine
Pittsburgh, PA 15260
E-mail: faeder@pitt.edu, Voice: (412) 648-8171, Fax: (412) 648-3163

Boris N. Kholodenko
Dept. of Pathology, Anatomy and Cell Biology
Thomas Jefferson University, JAH
1020 Locust Street
Philadelphia, PA 19107
E-mail: Boris.Kholodenko@jefferson.edu, Fax: (215) 923-2218
Abstract
The coupling of membrane-bound receptors to transcriptional regulators and other effector functions is mediated by multi-domain proteins that form complex assemblies. The modularity of protein interactions lends itself to a rule-based description, in which species and reactions are generated by rules that encode the necessary context for an interaction to occur, but also can produce a combinatorial explosion in the number of chemical species that make up the signaling network. We have shown previously that exact network reduction can be achieved using hierarchical control relationships between sites/domains on proteins to dissect multi-domain proteins into sets of non-interacting sites, allowing the replacement of each “full” (progenitor) protein with a set of derived auxiliary proteins. This decomposition into so-called macroscopic (macro) variables often greatly reduces network size. We describe here a method for automating domain-oriented model reduction and its implementation as a module in the BioNetGen modeling package, which takes as input a standard BioNetGen model and automatically performs the following steps: 1) Detecting the hierarchical control relationships between sites; 2) building up the auxiliary proteins; 3) generating a raw reduced model; and 4) cleaning up the raw model to provide the correct mass-balance for each chemical species in the reduced network. We tested the performance of this module on models representing portions of growth factor receptor and immunoreceptor-mediated signaling networks, and confirmed its ability to reduce the model size and simulation cost by at least one or two orders of magnitude. Limitations of the current algorithm include the inability to reduce models based on implicit site dependencies or heterodimerization, and loss of accuracy when dynamics are computed stochastically.
1 Introduction

1.1 Combinatorial complexity of cell signaling networks

Many signaling proteins, such as membrane receptors and their cytoplasmic adapters, have multi-domain structures and display multiple docking sites that engage several downstream signaling proteins, thereby serving as scaffolds [1-6]. In the conventional mechanistic framework, the functional states of such a multi-domain protein depends on the states of all domains of that protein. Each domain can assume multiple states, for instance, a docking site on a scaffold protein can be unphosphorylated and free, phosphorylated and free, phosphorylated and bound to a partner, which in turn can be unphosphorylated and free, or phosphorylated and bound to another protein or lipid, and so on.

As an example, we consider a cell-surface receptor of the receptor tyrosine kinase (RTK) family. A typical RTK has several domains and binding sites (see Fig. 1): the extracellular domain for ligand binding and receptor dimerization, and a cytoplasmic domain, which possesses tyrosine kinase activity and contains phosphorylation sites with tyrosine, serine and threonine residues. Following ligand binding, RTKs undergo receptor dimerization (e.g., epidermal growth factor (EGF) receptor) or an allosteric transition (e.g., insulin receptor, IR, and insulin-like growth factor receptor, IGF-1R) [7, 8]. These structural transitions result in the activation of intrinsic tyrosine kinase activity and subsequent autophosphorylation, which initiates signal processing through a battery of receptor interactions with adapter and target proteins containing characteristic protein domains, such as Src homology (SH2 and SH3), phosphotyrosine binding (PTB) and pleckstrin homology (PH) domains (reviewed in [7, 9, 10]). These proteins, in turn, can also posses multiple domains and sites that can be phosphorylated by the receptor and dephosphorylated by phosphatases.

Binding and phosphorylation/dephosphorylation reactions on distinct sites and diverse proteins that form protein complexes can occur sequentially or simultaneously. All these distinct possibilities multiply and generate the enormously large number of different micro-species (molecules or complexes of molecules in which each molecule is in distinct state) in a network. The concentration of each species is described by a separate ordinary differential equation (ODE), referred to as a chemical kinetic equation [11], leading to huge numbers of ODEs to describe even a few initial steps following the ligand binding to a receptor [12-15]. As a result, for a network model that includes multi-domain proteins and their complexes, simulations that
use a mechanistic description of all feasible chemical species and reactions rapidly become insurmountable even for the most powerful computers. For example, size estimates have been given recently for networks arising from the interactions of growth factor receptors and their initial scaffolding partners of $10^8$ [15] and $10^{23}$ [16].

Fig. 1. Multiplicity of the states of receptor and receptor-adaptep complexes. The state of the receptor molecule $R$ is characterized by a vector $(r_1, r_2, r_3, r_4)$, where $r_1$ stands for the ligand (L)-binding site, $r_2$ depicts the dimerization domain, and $r_3$ and $r_4$ specify the state of docking sites for adapter proteins. The adapter protein $B$ is a scaffold that possesses three sites (site $b_1$ for binding to the receptor, and tyrosine residues $b_2$ and $b_3$).

1.2 Domain-oriented model reduction

It has recently been shown that for many signaling networks a mechanistic picture of all possible species can be reduced to a description in terms of the experimentally measurable states of separate domains, referred to as macro variables [1, 2, 4, 17]. This domain-oriented reduction is based upon the mutual independence and hierarchical control relationships between different
sites of every protein in a network. If the rates of transitions between the states of site $q_i$ on a protein $Q$ depend upon the state of another site $q_j$ on the same protein $Q$, then site $q_i$ is termed *dependent* on site $q_j$, and, respectively, site $q_j$ is referred to as a *controlling* site for $q_i$ [1].

The independence of one site from another means that the time course of reactions involving one site, such as phosphorylation, dephosphorylation, and binding and release of downstream proteins etc., may be analyzed separately from the analysis of the entire scaffold. In fact, this approach will allow us to follow the fate of a particular site (or group of sites) on a scaffold independently of reactions occurring at other sites. Adding up the concentrations of all forms of the scaffold protein displaying a particular state of the site $b_i$ and all the sites that control the site $b_i$, we introduce macro variables that may be interpreted as concentrations of *auxiliary* proteins. Auxiliary proteins have fewer sites than the original scaffold (the *progenitor* protein).

The macro variables do not follow the fate of all species and reactions that are generated by scaffold signaling, thereby greatly reducing the number of states and equations required for a quantitative analysis of the system behavior. Macro variables also provide a direct connection to experimentally observable variables, because they usually can be quantified in experimental studies by Western blot analysis using site-specific antibodies, whereas micro-species of a scaffold protein cannot be assessed readily. Models transformed using macro variables can be simulated deterministically ODEs or stochastically using kinetic Monte Carlo methods, although simulations using the latter methods [2] require modifications to avoid loss of accuracy (see Sec. 2.5). For deterministic simulations, however, the transformed equations are exact in terms of the macro variables, so an exact model reduction is achieved provided one does not need the concentrations of micro-species, which is not a major restriction because these usually cannot be measured.

Although this domain-oriented approach to model reduction can decrease the number of variables by orders of magnitude, the manual definition of a set of macro variables for a highly interconnected network requires considerable modeling experience and effort and may obscure the basic structure of the model. There is an unmet need for the development of automatic procedures to liberate the user from manual derivation of kinetic equation systems in terms of macro variables.

**1.3 Rule-based model description**
Recent work has shown that automated generation of domain-oriented models can be accomplished through the use of a rule-based model description. Several software packages, such as Moleculizer [18], StochSim [19-21], BioNetGen [14, 22-24], all based on specifying multi-state proteins as software objects and applying transformation rules to generate the network of species and reactions arising from the rules. Rules represent a generalization of reactions, and a single rule may be applied to many different species to generate new reactions and new species as products. In the present paper, we will describe our domain-oriented reduction algorithm and examples using the specific syntax of the BioNetGen Language (BNGL), which is closely related to the κ-calculus of Danos and co-workers [16, 25], although the method could be applied to any domain-oriented model specification. We have also implemented the algorithm as a module of BioNetGen, which is freely available from http://bionetgen.org. A brief overview of BNGL is provided in the Appendix.

2 Results

2.1 Automatic construction of auxiliary proteins

The aim of the algorithm is to determine the set of auxiliary proteins containing subsets of progenitor sites to achieve maximum reduction in network size without loss in the accuracy of predicting macroscopic variables, which are defined by the user-specified reaction rules and observables. The algorithm begins by partitioning the sites on each molecule into (possibly overlapping) sets based on the control hierarchy. Redundant sets are then removed, and auxiliary proteins based on the controlling sets are introduced. These three steps are carried out as follows:

(1) Initial determination of controlling sets. For each protein (e.g., the protein named Q) with n sites \( q_1, \ldots, q_n \), we analyze the reaction rules as follows. If in the BNGL file, there is a reaction rule that includes the protein Q, in which site \( q_1 \) is present together with another site \( q_k \) and the state of site \( q_1 \) changes without the state of site \( q_k \) changing then site \( q_k \) is a controlling site for \( q_1 \). If \( q_1 \) and \( q_k \) change their states simultaneously in a reaction rule, or if \( q_1 \) and \( q_k \) are mentioned simultaneously in an observable, then sites \( q_1 \) and \( q_m \) are mutually dependent. Sites with identical names in the same molecule are also assumed to be mutually dependent in order to prevent dissection of proteins that can serve as a bridge for the formation of dimers. If two sites, \( q_1 \) and \( q_k \), are mentioned simultaneously in the pattern of an observable, then \( q_1 \) and \( q_k \) are also
mutually dependent. After finding all sites that control site \( q_1 \) as explained above, we repeat this procedure iteratively for each of the controlling sites found at the previous iteration until no new controlling sites can be found. The set that combines the site \( q_1 \) (by the definition, any site controls itself) and all direct or indirect controlling sites is termed a controlling set \( Z(Q,q_1) = (q_1,\ldots,q_s) \), \( s \leq n \), for site \( q_1 \) on \( Q \). Likewise, the sets \( Z(Q,q_i) \), \( i = 2,\ldots, n \), are determined for each remaining site on the protein \( Q \).

(2) Refinement of controlling sets. Controlling sets for different sites may overlap. Moreover, one controlling set (e.g., \( Z(Q,q_k) \) for site \( q_k \)) can be a subset of the controlling set for another site (e.g., \( Z(Q,q_i) \) for site \( q_i \)). To minimize the total number of auxiliary proteins that are introduced, we remove controlling sets that are subsets of other controlling sets because these will produce macro variables that contain no additional information. The controlling sets remaining after this refinement procedure are renumbered and designated as \( (Z_1(Q), \ldots, Z_m(Q), \ldots) \), \( m \leq n \). Note that after renumbering, we lose any information on the relationships between indexes 1, \ldots, \( m \) used for numbering sets \( Z_j \) and particular sites \( q_i \) on the protein \( Q \).

(3) Auxiliary protein definition. For each set \( Z_j(Q) \), \( j = 1,\ldots, m \), we associate the function \( Q_j(Z_j(Q)) \) derived as the sum of the concentration of protein \( Q(q_1,\ldots,q_n) \) over all possible states of the sites that are not included in set \( Z_j(Q) \). Consequently, \( Q_j \) depends on the states of the sites that belong to \( Z_j(Q) \) but is independent of all other sites that do not belong to \( Z_j(Q) \). To transform the rule set defining the model from micro- into macro-variables, we define for each macro function \( Q_j \) an auxiliary protein \( Q_j \) having a set of sites \( (Z_j(Q)) \), which is a subset of the domains on the progenitor protein \( Q \). In physical terms, the multi-state progenitor protein is replaced by a number of auxiliary proteins, each with a smaller number of sites.

We can illustrate this procedure for the simple example of proteins \( R \) and \( B \) shown in Fig. 1. Analysis of the reaction rules that describe binding and phosphorylation reactions that involve \( R \) and \( B \) (see Supplement 1) shows that on \( R \) phosphorylation residues, \( r_3 \) and \( r_4 \), depend on the ligand-binding site, \( r_1 \), as well as on the dimerization site, \( r_2 \). Likewise, on the scaffolding adapter protein \( B \), the RTK-binding site \( b_1 \) controls the phosphorylation residues \( b_2 \) and \( b_3 \).

Controlling sets of the sites on \( R \) and \( B \) are determined by the algorithm described above as follows, \( Z(R,r_1) = \{r_1\} \), \( Z(R,r_2) = \{r_1,r_2\} \), \( Z(R,r_3) = \{r_1,r_2,r_3\} \), \( Z(R,r_4) = \{r_1,r_2,r_4\} \), \( Z(B,b_1) = \{b_1\} \), \( Z(B,b_2) = \{b_1,b_2\} \), \( Z(B,b_3) = \{b_1,b_3\} \). The deletion of redundant sets results in the following remaining sets, \( Z_1(R) = \{r_1,r_2,r_3\} \), \( Z_2(R) = \{r_1,r_2,r_4\} \), \( Z_1(B) = \{b_1,b_2\} \), \( Z_2(B) = \{b_1,b_3\} \).
Although it may first appear counterintuitive, the extent of model compression increases with the number of the auxiliary proteins derived from each protein $Q$, since the total number of micro-states is a product of the number of states of each site on $Q$, whereas the number of macro-states of is a sum of the number of states of each auxiliary protein $Q_j$. In the extreme case of interactions among all sites on the scaffold, the above procedure results in a single controlling set that contains every site on the protein. The resulting single auxiliary protein $Q_t = Q(q_1,\ldots,q_n)$ is then the same as the progenitor protein, and no model reduction occurs.[26]

2.2 Generation of reactions and observables that preserve mass-balance

Sites found on more than one auxiliary protein derived from the same progenitor protein are termed shared sites. If a particular site is found on only one auxiliary protein, this site is referred to as a unique site. For instance, sites $r_1$ and $r_2$ on the RTK $R$ and the $b_1$ on the adapter $B$ in Fig. 1 are shared, whereas $r_3$ and $r_4$, $b_2$ and $b_3$ are unique. Care must be taken to ensure that the total mass of proteins that bind to shared sites is conserved. Without such care, the introduction of $n$ auxiliary proteins containing the same shared site leads to an $n$-fold increase in the concentration of the shared site and produces incorrect binding kinetics. As shown previously [1, 2], the correct kinetics is obtained if only one of the binding reactions involving the shared site consumes or produces the binding partner. The auxiliary protein involved in this reaction is termed balance-accountable, whereas the remaining auxiliary proteins are termed balance-unaccountable. The choice of the balance-accountable auxiliary protein among the auxiliary proteins containing the shared site is arbitrary [1]. A detailed example that illustrates how this may be implemented in BioNetGen using non-consumption tags and a manually specified macro reduction is provided in Supplement 2.

The procedure described above is insufficient when both reactants in a binding reaction contain shared sites. This is an important case to consider because many, if not most, RTKs dimerize. For this reason the implementation of the macro module in BioNetGen carries out mass-balance corrections in a different way that does not involve the use of non-consumption tags in reaction rules, but rather applies corrections to the network of species and reactions generated by rule application, i.e., at the level of the net-file rather than at the level of the bngl-file (see Appendix). Detailed description of the implemented procedure is provided in
Supplement 3, but the essential elements comprise steps 4(a)-4(c) in the algorithm summary provided below.

### 2.3 Summary of domain-oriented reduction algorithm

![Flowchart of operations for the domain-oriented reduction algorithm](image)

**Fig. 2. A flowchart of operations for the domain-oriented reduction algorithm.**

Summarizing the algorithms described above, the module for domain-oriented model reduction applies the following operations to produce a correct reduced network model (see Fig. 2):

1. Analysis of reaction rules and patterns of the observables to determine the site dependence hierarchy for each protein.
2. Replacement, where applicable, of progenitor proteins with the sets of auxiliary proteins.
3. Generation of “raw” or uncorrected network of species and reactions (accomplished in BioNetGen by the `generate_network` command).
4. Correction of the raw macro-network model.
   a. Complexes that contain different auxiliary proteins derived from the same progenitor protein are removed from the list of species and the lists of species referenced by each observable. Reactions involving the removed species are also eliminated.
   b. Reactions are corrected to eliminate consumption or production of species that bind to or dissociate from shared sites of balance-unaccountable
auxiliary proteins. An exception is made for the case in which an auxiliary protein dimerizes.

c. Observables are corrected to eliminate species that contain balance-unaccountable auxiliary proteins if their contribution to the observable has been also taken into account by species that contain balance-accountable proteins.

2.4 Numerical examples

Numerical experiments illustrate the performance of automated model reduction methods for a set of several RTK signaling networks, including an EGFR-like network, in which ligand binding induces aggregation through receptor-receptor interactions [17, 27, 28] (see Fig. 1 and Supplement 1), and an FcεRI-like network (see Supplement 4), in which receptor aggregation is mediated by a bivalent ligand [29]. There are two versions of both models, one with two receptor tyrosine residues, \( r_3 \) and \( r_4 \), which upon phosphorylation can bind the adapter proteins, \( A \) and \( B \), respectively (see Fig. 1), and one with an additional tyrosine, \( r_{3a} \), which also can bind \( A \) upon phosphorylation.

Table 1 shows the extent of model reduction provided by the domain-oriented method. Although the models presented here are rather small, including only four proteins and a few reaction rules, the macro-reduction method decreases the number of species and reactions, as well as time required for model generation, by orders of magnitude. Since even the reduced (macro-) models contain tens, if not hundreds of species, and hundreds of reactions, manual (non-automatic) preparation of the macro-models seems impractical. The difference between results for the full (micro-) and reduced (macro-) models for the computed values macroscopic observables is less than \( 10^{-8} \) (the tolerance limit for the ODE integration), which confirms the correct performance of the macro-reduction algorithm (data not shown).
Table 1. Quantification of the network reduction achieved by the domain-oriented (macro) reduction method

<table>
<thead>
<tr>
<th>Network</th>
<th>Species in the micro-/macro-model</th>
<th>Reactions in the micro-/macro-model</th>
<th>Network Generation CPU time (s)</th>
<th>ODE Integration CPU time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EGFR-like network</td>
<td>Receptor with 2 tyrosine residues</td>
<td>708/108</td>
<td>7432/534</td>
<td>51.6/8.45</td>
</tr>
<tr>
<td></td>
<td>Receptor with 3 tyrosine residues</td>
<td>6000/135</td>
<td>81364/642</td>
<td>662.0/12.0</td>
</tr>
<tr>
<td>FcεRI-like network</td>
<td>Receptor with 2 tyrosine residues</td>
<td>213/48</td>
<td>2230/198</td>
<td>14.2/3.87</td>
</tr>
<tr>
<td></td>
<td>Receptor with 3 tyrosine residues</td>
<td>1599/60</td>
<td>22990/240</td>
<td>182.4/6.02</td>
</tr>
</tbody>
</table>

1Computed using BioNetGen 2.0.41 running on Pentium® 4 CPU 2.80 GHz with 1 GB RAM.

2.5 Limitations

Although these examples confirm the ability of the module to reduce the models by at least one or two orders of magnitude, several aspects of the algorithm give rise to limitations, which are summarized here and described in detail in Supplement 5, where extensions of the algorithm to address these limitations are also proposed.
In our current algorithm, each control relationship is detected using a single reaction rule that is taken separately from other rules. As a result, the algorithm cannot identify control relationships arising from implicit binding relationships, such as in the BNGL expression A . B, which requires that proteins A and B be in the same complex but does not specify the mechanism of binding. This limitation can be addressed at the user level by avoiding implicit dependencies in the model specification, although cases arise when this is not possible (see, e.g. [30]). As detailed in Supplement 5, iterative processing of the rules could be used to resolve these control relationships. Similarly, possible chains or loops of chemical bonds that link a unique site of an auxiliary protein to the shared sites of the same protein or another instance of the same protein type are not detectable using our single-pass algorithm. Handling of such cases would also require iterative processing of the rules.

The algorithm presented here only allows detection of hierarchical control relationships within a single protein. Models that allow multiple interactions between two proteins may give rise to control relationships that span multiple proteins, and, in such cases the single protein analysis used here gives an incorrect reduced model. An example involving ordered phosphorylation of an adapter protein is illustrated in Fig. S5.2 and a bng1-file for the model is given in Supplement 6. Detection of such complex control hierarchies is not currently implemented, but the user may prevent incorrect model reduction from occurring by manually disabling macro-reduction of problematic proteins. In addition, incorrect model reduction can be detected by comparing time courses of the reduced and exact models, as shown in Fig. S5.3.

The most serious drawback of the current algorithm is that it does not find identify reductions for the case of heterodimerization, in which auxiliary proteins from different progenitor proteins binding to each other via shared sites (either directly or indirectly). This is an important case for biological models to consider because many signaling proteins form such complexes, a prominent example being the ErbB family of RTKs [31, 32]. Allowing association of each auxiliary protein with all auxiliary proteins derived from every other progenitor protein dramatically increases the number of macro-species. Due to possible different numbers of auxiliary proteins derived from each progenitor protein that undergo heterodimerization, this problem does have the elegant solution that homodimerization has (when the complexes with different auxiliary proteins derived from the same progenitor protein are simply removed from the model).
A final limitation that applies to the BioNetGen implementation but not to the reduction algorithm per se is that simulations using kinetic Monte Carlo methods such as Gillespie’s algorithm [33] with the macro-reduced reaction network will not be exact unless reactions involving binding and dissociation of shared sites are properly correlated. The problem arises because in a discrete-event simulation, every time binding or dissociation event occurs involving a shared site, the event should apply to all of the shared sites of the same molecule. This can result in different levels of shared site occupancy for each of the auxiliary proteins sharing the site, which is clearly non-physical. (This problem does not apply to the ODE equations because all the events occur at the same rate and thus give the same values of site occupancy.) To retain the correct site occupancies, one could apply correlated Monte Carlo sampling [34], in which one event is used to trigger a change in state of the shared site of all \( n \) auxiliary proteins. This has not been done for the stochastic simulation algorithm in BioNetGen, but could be easily encoded in models exported in the Systems Biology Markup Language (see Appendix).

3 Discussion

The multiplicity of scaffold proteins involved in RTK signaling networks, their sites and states of these sites results in a combinatorial explosion of the number of possible states that involved proteins and their complexes may have. The interactions present in signal transduction systems may easily imply networks of possible species and reactions that are too large to simulate using standard methods for chemical kinetics. Recently, advances in kinetic Monte Carlo methodology that use particle-based event-driven simulations to avoid explicit generation of species and reactions appear to have broken to bottleneck caused by combinatorial complexity [35, 36]. ODEs, however, afford both computational and analytical advantages over stochastic methods and therefore methods for limiting the size of the ODE system implied by a set of biochemical interactions will continue to be important.

A central result of our previous theoretical studies is that for many signaling networks, a mechanistic picture of all possible species may be substituted with a more compact model that describes the network in terms of experimentally detectable states of separate domains [1, 2, 4, 17]. The key features that allow such domain-oriented reduction are hierarchical control relationships between sites on proteins involved in signaling networks.
Based on these findings, we have developed a method for automatic domain-oriented reduction of signaling network models, which is implemented as a module in the software package BioNetGen. An input bngl-file is preprocessed by the new reduction module. First, the module determines the control relationships between sites on protein molecules. Second, if possible, self-controlling subsets of sites are determined for each protein, and each reducible protein (progenitor protein) is substituted with a set of auxiliary proteins that have only the sites that belong to the self-controlling subsets. Third, the raw network model, which is described in terms of auxiliary proteins, is generated using BioNetGen. Finally, the raw model is corrected to provide correct mass-balance for each species in the reduced model. The algorithm has been applied to several realistic examples involving aggregation of receptors with multiple binding and modification sites, and a high degree of model reduction was achieved, resulting in several orders of magnitude of increased computational efficiency with no loss of accuracy. The method is fully automated and does not require the user to understand details of the algorithm. Future work will focus on overcoming limitations to the applicability of the algorithm outlined in Sec. 2.5 and detailed in Supplement 5. Recently, a new model reduction technique based on modular analysis has been proposed that augments the domain-oriented approach used here, increasing the level of compression that can be attained at the cost of introducing some degree of error, which appears to be small for the cases examined so far [15]. At the present time, the method requires manual analysis and application, but its automation would appear to be a promising area for future development.

4 Appendix: Overview of BioNetGen

BioNetGen provides a flexible language for the description of protein structure and protein interactions called the BioNetGen language (BNGL) [37]. The model specification in the bngl-file consists of four elements: parameters, species, reaction rules, and observables. Parameters specify the kinetic rate constants, total protein concentrations and other fixed numerical properties of the model. Species describe molecules that are present at the start of network generation. For example, the species \( P(s_1, s_2-pY) \) defines a protein named \( P \), which has two sites named \( s_1 \) and \( s_2 \), and specifies that the site \( s_1 \) is free, and the site \( s_2 \)
is phosphorylated and also free. Reaction rules list the rules for building the biochemical network. For example, the reversible rule

\[ A(b) + B(a,c) \rightleftharpoons A(b!1).B(a!1,c) \, k_{on}, \, k_{off} \]

describes the binding and dissociation of molecules \(A\) and \(B\), where the first reactant may be any species that contains the protein \(A\) whose site \(b\) is free, and the second reactant may be any species that contains the protein \(B\) whose sites \(a\) and \(c\) are both free. The product of these reactions contains proteins \(A\) and \(B\) bound via \(b\)-site on \(A\) and \(a\)-site on \(B\), as indicated by the exclamation mark followed by the number 1, which denotes a termination point for the bond labeled ‘1’. In this complex, the \(c\)-site on \(B\) is free and all other sites on \(A\) or \(B\) (that were specified in the species block) may be in any possible state. All the binding reactions generated by this reaction rule will have a second-order rate constant \(k_{on}\), and all dissociation reactions will have the first-order rate constant \(k_{off}\). Observables describe the sums over the concentrations of species sharing similar attributes, which correspond to the quantities that are measured in typical biological experiments. For example, the observable

\[ \text{Molecules } P\_s2\_phos \, P(s2\_pY) \]

defines the observable named \(P\_s2\_phos\) of type \(\text{Molecules}\), which means a weighted sum over the species matching the pattern \(P(s2\_pY)\), which finds instances of the protein \(P\) in which the site \(s_2\) is in state \(pY\).

The command \texttt{generate_network} in BioNetGen2 automatically generates a complete biochemical network model that contains all feasible species and reactions by iterative application of the rules to the initial set of species. The resulting network can be written either in the BIONETGEN-specific format (\texttt{net-file}) or exported in the Systems Biology Markup Language [38], which can be imported by a large number of other simulation and analysis tools. In both cases the generated network can be simulated using either the ODE or stochastic approach.

5 Acknowledgements

The authors would like to thank Dr. Mikhail V. Kravchenko from the State Research Center – Institute of Biophysics, who substantively rewrote, rearranged and optimized the Perl code for the whole domain-oriented reduction module. BNK acknowledges support from NIH grants R01-
GM059570 and R33-HL088283 (NHLBI Exploratory Program in Systems Biology). JRF acknowledges support from NIH grants R37-GM35556 and R01-GM076570. A portion of this work was completed during JRF’s tenure as Technical Staff Member at Los Alamos National Laboratory. Participation of the State Research Center – Institute of Biophysics was facilitated by the Civilian Research and Development Foundation, Grant 1624.

6 References


26. Note that although in the paper (both in the Main text and in the Supplements), for the sake of simplicity, we label the auxiliary proteins by index, e.g., Q1, Q2, Q3, the program module for macro-reduction labels auxiliary proteins using its contained sites (e.g., Q__q1_q2(q1,q2), Q__q1_q3(q1,q3) etc.).
33. Gillespie, D.T., 'A general method for numerically simulating the stochastic time
35. Danos, V., J. Feret, W. Fontana, and J. Krivine, 'Scalable Simulation of Cellular
    for Rule-based Modeling of Biochemical Networks'. submitted.
    Bornstein, D. Bray, A. Cornish-Bowden, A.A. Cuellar, S. Dronov, E.D. Gilles, M.
    Ginkel, V. Gor, I.I. Goryanin, W.J. Hedley, T.C. Hodgman, J.H. Hofmeyr, P.J. Hunter,
    Lucio, P. Mendes, E. Minch, E.D. Mjolsness, Y. Nakayama, M.R. Nelson, P.F. Nielsen,
    language (SBML): a medium for representation and exchange of biochemical network