Published in IET Systems Biology
Received on 19th December 2007
Revised on 8th June 2008
doi: 10.1049/iet-syb:20070081

Special Issue – Selected papers from the First q-bio Conference on Cellular Information Processing

ISSN 1751-8849

Domain-oriented reduction of rule-based network models

N.M. Borisov[1,2] *A.S. Chistopolsky*[2] *J.R. Faeder*[3] *B.N. Kholodenko*[1]

[1]*Department of Pathology, Anatomy and Cell Biology, Thomas Jefferson University, 1020 Locust St., Philadelphia, PA 19107, USA*
[2]*Burnasyan Federal Medical Biophysical Center, 46, Zhivopisnaya St., Moscow 123182, Russia*
[3]*Department of Computational Biology, University of Pittsburgh School of Medicine, Pittsburgh PA 15260, USA*
E-mail: faeder@pitt.edu

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 signalling network. The authors 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 (offspring) proteins. The description of a network in terms of auxiliary proteins that have fewer sites than progenitor proteins often greatly reduces network size. The authors describe here a method for automating domain-oriented model reduction and its implementation as a module in the BioNetGen modelling package. It 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. The authors tested the performance of this module on models representing portions of growth factor receptor and immunoreceptor-mediated signalling 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 heterodimerisation and loss of accuracy when dynamics are computed stochastically.

1 Introduction

1.1 Combinatorial complexity of cell signalling networks

Many signalling proteins, such as membrane receptors and their cytoplasmic adapters, have multi-domain structures and display multiple docking sites that engage several downstream signalling proteins, thereby serving as scaffolds [1–6]. 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. In general, the functional states of such multi-domain proteins will depend on the states of all domains of the protein. We define a microscopic model as one that explicitly represents all possible states of multi-domain proteins and the feasible reactions among these states.

As an example, we consider a cell-surface receptor of the receptor tyrosine kinase (RTK) family. RTK's have a modular structure that can be divided into an extracellular region, which contains the ligand-binding and receptor dimerisation sites and a cytoplasmic region, which has tyrosine kinase activity and contains phosphorylation sites with tyrosine, serine and threonine residues (Fig. 1). Ligand binding activates RTKs by inducing either dimer formation (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 receptor interactions with a battery of 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 possess multiple domains and sites that can be phosphorylated by the receptor and dephosphorylated by phosphatases.

Binding between two signal-transduction proteins often requires one of the two interacting sites to be phosphorylated, which imposes an ordering on phosphorylation and binding events. For proteins that have multiple binding sites, however, binding of other proteins at different sites may be independent (i.e. no interaction among binding partners) or cooperative (i.e. binding partners interact either positively or negatively). Ordering is imposed on binding interactions at two different sites only if the cooperativity is

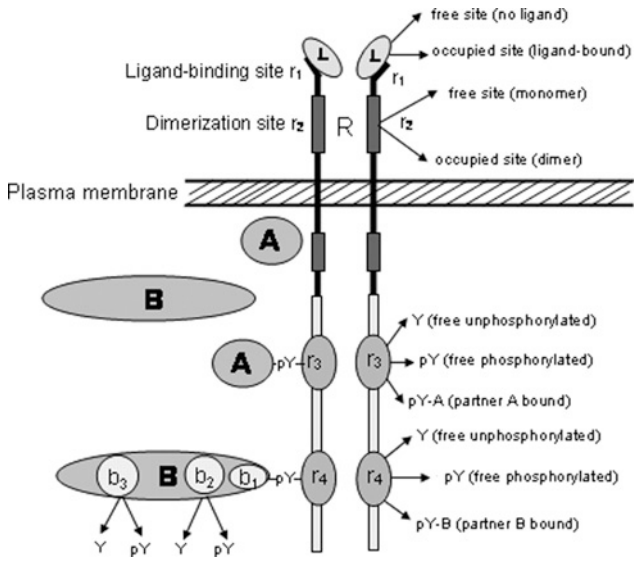

Figure 1 *Multiplicity of the states of receptor and receptor–adapter complexes*

State of the receptor molecule R is characterised 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 dimerisation site and r_3 and r_4 specify the state of docking sites for adapter proteins

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)

complete, that is, one site must be occupied for binding to occur at the second site or one site must not be occupied for binding to occur at a second site. Thus, in most cases the presence of multiple binding sites gives rise to many different combinations of protein aggregates that can have a large number of different functional states. In general, the number of states of an aggregate grows in a multiplicative fashion with the number of possible states of each site, leading to a combinatorial explosion in the total number of different chemical species (molecules or complexes of molecules in which each molecule is in a distinct state) that must be included within a microscopic model. In the conventional approach to modelling chemical kinetics [11], in which the concentration of each species is described by a separate ordinary differential equation (ODE), combinatorial complexity may generate such a large set of equations that simulation becomes infeasible on even the most powerful computers. The problem arises in models describing only the initial steps following ligand-receptor binding, which can lead to hundreds or thousands of possible species [12–15]. Extended models of growth factor receptors and their initial scaffolding partners can produce networks of 10^8 species [15], 10^{23} species [16], and beyond, rendering the conventional approach useless for such microscopic models.

1.2 Domain-oriented model reduction

One way to avoid the problem of combinatorial explosion is model reduction. It has recently been shown that by introducing a set of variables that tracks only subsets of the possible combinations of the domain/site states rather than the full set of possible complexes, it is possible to derive a reduced set of dynamical equations for many signalling networks [1, 2, 4, 15, 17]. This domain-oriented approach to model reduction is based upon the mutual independence and hierarchical control relationships between different sites of each protein in a network, which goes as follows. 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 sites means that the time course of reactions involving some sites may be decoupled from the reactions occurring at other sites. For each scaffold protein, called a progenitor, a set of auxiliary (offspring) proteins can be introduced, each of which contains a subset of the progenitor protein's sites. Previous work has shown that the sites contained by the auxiliary proteins can be chosen so that each reacts independently of the other auxiliary proteins. The concentration of an auxiliary protein with sites $q_1 \ldots q_k$ in states $s_1 \ldots s_k$ is defined to be the sum of concentrations of all forms of the scaffold protein in which each of the k sites has the same state as in the auxiliary protein. The concentrations of the auxiliary proteins are thus macroscopic (macro) variables that are comprised of sums over the concentrations of microscopic (micro) species in the system. In contrast to the number of

micro variables, which is a multiplicative function of the number of states of each site, the number of macro variables is additive in the number of states of each auxiliary protein. If a protein contains multiple independent sites, the number of macro variables describing the protein's dynamics can be much smaller than the number of micro states of the protein.

The domain-oriented approach thus provides a macroscopic description of network dynamics in that it does not follow the fate of all species and reactions that are generated by scaffold signalling, thereby greatly reducing the number of states and equations required for a quantitative analysis of the system behaviour. The ODE's obtained by the transformation to macro variables are exact in terms of auxiliary proteins. Kinetic Monte Carlo methods, such as the Gillespie algorithm [2], can also be used to provide an exact stochastic description of the dynamics in terms of the macro variables, but, as we note below in Section 2.4 require slight modification to avoid loss of accuracy. The transformation to macro variables entails some loss of information about correlations between independent sites of a protein, but such correlations typically cannot be measured by available experimental techniques, most of which detect binding or phosphorylation at either the whole protein or the single site level. If such data is available, the modeler may choose to define observables that track multiple sites within a protein, although this will lessen the extent to which the model can be reduced. Multi-site observables may also be approximately reconstructed from single-site observables [1, 2, 4]. In practice, single-site macro variables are frequently sufficient for making direct comparisons with experimental measurements [18].

The domain-oriented approach to model reduction can decrease the number of variables by orders of magnitude and thus promises to be a powerful tool for the development of realistic models of intracellular signalling. Existing methods [1, 2, 4, 15, 17], however, are not automated and require the modeller to define manually the set of auxiliary proteins and the rules for their interaction. For a highly interconnected network, this requires considerable modelling experience and effort and may obscure the basic structure of the model. Moreover, the procedure has not yet been formalised algorithmically in the previous works that describe the principle of domain-oriented reduction. In this paper, we present an algorithm for domain-oriented model reduction, which has been implemented as module in the freely-available BioNetGen modelling package [19].

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 [20], StochSim [21–23], BioNetGen [14, 18, 19, 24], enable the develop of rule-based models based on a multi-state description of proteins and other signalling molecules and rules that transform these molecules according to specified properties of the reactants. Rules represent a generalisation of reactions, and a single rule may be applied to many different species to generate new reactions and new species as products. In order to simulate a rule-based model as a set of ODEs, rules are applied iteratively to a seed set of species to generate all of the possible reactions and species in the network [24]. The cost of network generation, as well as subsequent ODE integration, can be become prohibitive for models exhibiting a high degree of combinatorial complexity. The goal of the current algorithm is to reduce the costs of network generation and simulation by replacing each multi-state progenitor protein in the model with a set of derived auxiliary (offspring) proteins that group sets of independently-acting sites. Application of the transformed rules to the set of auxiliary proteins will then generate a transformed network that is smaller in size but no less accurate for predicting the time evolution of the macro variables.

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 with further details provided in [19].

2 Results

Domain-oriented reduction attempts to construct the smallest possible model of the network given the molecules and interactions specified by the user that still allows correct calculation for the specified observables. Since the domain-oriented reduction method relies on control relationships between protein sites, the module should perform at least two major actions:

(1) Construct auxiliary proteins by detecting control relationships between progenitor protein sites.

(2) Generate reactions and observables for the reduced network that preserve mass balance.

2.1 Automatic construction of auxiliary proteins

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:

IET Syst. Biol., 2008, Vol. 2, No. 5, pp. 342–351
doi: 10.1049/iet-syb:20070081

(1) Initial determination of controlling sets: The aim of this step is to determine the controlling set for each site on each protein according to the reaction rules and observable patterns specified in the `bngl`-file. For each protein Q with n sites called $q_1, \ldots, q_n$, we analyse the reaction rules as follows. If there is a reaction rule in which site q_1 of Q is present together with another site q_k and the state of site q_1 changes while the state of site q_k does not, 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_k are mutually dependent, which means that q_k is considered a controlling site of q_1 and vice versa. 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. After finding all sites that control site q_1, 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 $\mathbf{Z}(Q,q_1) = (q_1, \ldots, q_s)$ for site q_1 on Q. Likewise, the sets $\mathbf{Z}(Q,q_i)$, $i = 2, \ldots, n$, are determined for each remaining site on the protein Q. All controlling sets $\mathbf{Z}(Q,q_i)$ are subsets of the full set $(q_1, \ldots, q_n)$ of sites for protein Q. The resulting sets $\mathbf{Z}(Q,q_i)$ do not depend on the order of appearance of any protein, site, reaction rule, or observable in the `bngl`-file, because the process terminates only after all possible control relationships have been found.

(2) Refinement of controlling sets: The aim of the refinement step is to eliminate redundancy among the sets of controlling sets that are used to define the auxiliary proteins. Controlling sets for different sites may overlap, and if one controlling set is a subset of another it is removed from the set of controlling sets for a given progenitor, because an auxiliary protein defined from this redundant set would contain no unique information. The controlling sets remaining after this refinement procedure are renumbered and designated as $(\mathbf{Z}_1(Q), \ldots, \mathbf{Z}_m(Q), m \leq n)$. Note that after renumbering, we lose any information on the relationships between indexes $1, \ldots, m$ used for numbering sets $\mathbf{Z}_j$ and particular sites q_i on the protein Q. The refined set of controlling sets is optimal as the starting point for auxiliary protein definition because it is the smallest set of controlling sets for the sites of Q that contains all sites of Q.

(3) Auxiliary protein definition: For each set $\mathbf{Z}_j(Q)$, $j = 1, \ldots, m$, we define the macro variable $[Q_j(\mathbf{Z}_j(Q))]$, which is 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 $\mathbf{Z}_j(Q)$. For example, if $\mathbf{Z}_j(Q)$ contains all sites $(q_1, \ldots, q_n)$ except q_x, q_y, q_z, the corresponding variable is $[Q_j(\mathbf{Z}_j(Q))] = \sum_{q_x=0}^{X} \sum_{q_y=0}^{Y} \sum_{q_z=0}^{Z} [Q(q_1, \ldots, q_x, \ldots, q_y, \ldots, q_z, \ldots, q_n)]$, where indices q_x, q_y and q_z run over all the possible states (denoted form 0 to Z, Y and Z, respectively) of the sites q_x, q_y, q_z and $[Q(q_1, \ldots, q_x, \ldots, q_y, \ldots, q_z, \ldots, q_n)]]$ is the concentration of protein Q in the state $(q_1, \ldots, q_x, \ldots, q_y, \ldots, q_z, \ldots, q_n)$. Hence, the macro variable $[Q_j]$ depends on the states of the sites that belong to $\mathbf{Z}_j(Q)$ but is independent of all other sites that do not belong to $\mathbf{Z}_j(Q)$. To transform the rule set defining the model from the micro variables into the macro variables, we define an auxiliary protein Q_j for each macro variable $[Q_j]$. The auxiliary protein Q_j has a set of sites $(\mathbf{Z}_j(Q))$, 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 (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 dimerisation 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, $\mathbf{Z}(R,r_1) = \{r_1\}$, $\mathbf{Z}(R,r_2) = \{r_1,r_2\}$, $\mathbf{Z}(R,r_3) = \{r_1,r_2,r_3\}$, $\mathbf{Z}(R,r_4) = \{r_1,r_2,r_4\}$, $\mathbf{Z}(B,b_1) = \{b_1\}$, $\mathbf{Z}(B,b_2) = \{b_1,b_2\}$, $\mathbf{Z}(B,b_3) = \{b_1,b_3\}$. The deletion of redundant sets results in the following remaining sets, $\mathbf{Z}_1(R) = \{r_1,r_2,r_3\}$, $\mathbf{Z}_2(R) = \{r_1,r_2,r_4\}$, $\mathbf{Z}_1(B) = \{b_1,b_2\}$, $\mathbf{Z}_2(B) = \{b_1,b_3\}$.

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 variables is a product of the number of states of each site on Q, whereas the number of macro variables 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 $Q_1 = Q(q_1, \ldots, q_n)$ is then the same as the progenitor protein, and no model reduction occurs. 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, for example, Q1, Q2, Q3, the macro-program module labels auxiliary proteins using its contained sites [e.g. Q_q1_q2(q1,q2), Q_q1_q3(q1,q3), etc.].

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. The model reduction algorithm must ensure that proteins that bind to

shared sites will not be counted more than once in mass-balance equations. Otherwise, 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 done manually in BioNetGen scripts using non-consumption tags and a manually-specified macro reduction is provided in Supplement 2.

This procedure, however, 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 dimerise. For this reason the current domain-oriented reduction module for BioNetGen performs 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, that is, at the level of the `net`-file rather than at the level of the `bngl`-file (Appendix).

A 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.

(1) Analysis of reaction rules and patterns of the observables to determine the site dependence hierarchy for each protein, according to the algorithm described in Section 2.1

(2) Replacement, where applicable, of progenitor proteins with the sets of auxiliary proteins, according to the algorithm described in Section 2.1.

(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 two or more different auxiliary proteins derived from the same progenitor protein overload the macro-network with extra species. These species lead to the multiplication of the concentrations of unique sites, which leads to spurious effects. To eliminate this problem, all complexes that contain different auxiliary proteins derived from the same progenitor protein are removed from the list of species in the network and from the list of species corresponding to each observable. Reactions involving the removed species are also removed.

b. To provide the correct mass balance for the partners of the shared sites, the module disables consumption or production of all species that bind to or dissociate from shared sites of balance-unaccountable auxiliary proteins. The current version of the domain-oriented reduction module treats homo-dimerisation as an exception to this rule, but does not handle the case of binding between shared sites (either direct or mediated via other proteins) of different progenitor proteins (Section 2.4 and Supplement 5 for more detail).

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.

A flowchart of the complete algorithm for domain-oriented model reduction that is implemented as a BioNetGen module is shown in Fig. 2.

2.3 Numerical examples

Numerical experiments illustrate the performance of the automated model reduction methods for a set of several RTK signalling networks, including an EGFR-like network, in which ligand binding induces aggregation through receptor-receptor interactions [17, 26, 27] (Fig. 1 and Supplement 1) and an FcεRI-like network (Supplement 4), in which receptor aggregation is mediated by a bivalent ligand [28]. 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 (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 achieved by the domain-oriented method. Although the models

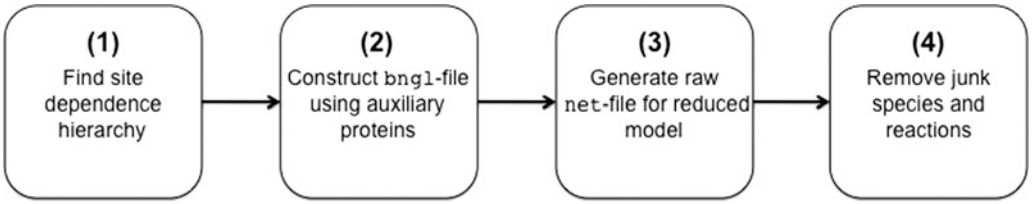

Figure 2 *Flowchart of operations for the domain-oriented reduction algorithm*

doi: 10.1049/iet-syb:20070081

presented here are small in scope, including only four proteins and a few reaction rules, the reduction method decreases the number of species and reactions, as well as time required for model generation, by orders of magnitude. Because even the reduced models contain tens, if not hundreds of species, and hundreds of reactions, manual (non-automatic) preparation of the reduced models seems impractical. The relative difference between results for the full and reduced models for the computed values of observables is less than 10^{-8} (the tolerance limit for the ODE integration), which confirms that the algorithm performs correctly and does not introduce significant numerical errors into the integration (data not shown).

2.4 Limitations

Although these examples confirm the ability of the module to reduce the models by at least one or two orders of magnitude, the algorithm has limitations, which are summarised here and described in detail in Supplement 5, where future extensions of the algorithm to address these limitations are also proposed. For each of the six cases discussed below, the possibility exists that current module may either fail to reduce a reducible model or may produce an incorrect reduced model (i.e. one whose simulation produces results that differ from those produced by the full model) if the model possesses certain features that trigger limitations in the current algorithm. To help users of the module avoid these outcomes, we provide tips, summarised in Table 2, for recognising problematic model elements and adjusting module control parameters to avoid reducing parts of a model that cannot be correctly handled. We strongly suggest that, wherever possible, simulation results from reduced models obtained by the macro module be compared with results of an exact simulation to verify that the model has been correctly reduced. Although the limitations described here and elaborated in Supplements 3 and 5 represent all limitations in the current algorithm of which we are presently aware, we do not have a proof that these are exhaustive, and it is thus possible that unforeseen instances of incorrect model reduction may occur.

1. Identical site names: As mentioned in Section 2.1, the module assumes mutual dependence among sites with identical names. Although this feature is necessary for the proper treatment of 'bridging' events, in which two monomers are linked by a bivalent ligand, it also decreases the extent of model reduction when such bridging is not necessary. In the latter case, the user is advised to use unique names for each site on a molecule.

2. Implicit bonds: In the 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 [29]. As detailed in Section 3 of Supplement 5, iterative processing of the rules could be used to resolve these control relationships.

3. Binding between shared and unique sites of the same auxiliary protein: The current module incorrectly reduces models generating complexes with chains or loops of chemical bonds that link a unique site of an auxiliary protein to a shared site of the same protein or another instance of the same protein type. The resulting reduced models have incorrect mass balances leading to incorrect simulation results for some observables. Automated handling of such cases would also require iterative processing of the rules. To avoid the possibility of an error in model reduction, the user is advised to validate results of reduced model against full model wherever possible. If a discrepancy occurs, the user can inspect the species list for the occurrence of complexes that link shared and unique

Table 1 Quantification of the network reduction achieved by the domain-oriented reduction method[a]

		Total number of species in the full/ reduced model	Total number of reactions in the full/ educed model	CPU time for network generation, of the full/reduced model, s	CPU time for ODE integration of the full/ reduced model, s
EGFR-like network	receptor with two tyrosine residues	708/108	7432/534	51.6/8.45	1.45/0.22
	receptor with three tyrosine residues	6000/135	81 364/642	662.0/12.0	12.58/0.76
FcεRI-like network	receptor with two tyrosine residues	213/48	2230/198	14.2/3.87	0.47/0.15
	receptor with three tyrosine residues	1599/60	22 990/240	182.4/6.02	3.58/0.12

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

Table 2 Overview of limitations in the current version of the domain-oriented reduction module

Limitation	Model properties that trigger this limitation	How the module processes this case	How to avoid the problem	Plans for future handling
1. identical site names	proteins containing two or more sites that have the same name	sites with identical names are assumed mutually dependent	make site names unique	none
2. implicit bonds	`include/exclude` directives in reaction rules. Implicit bonds in reaction rules or observables	implicit bonds are not considered in control relationships	remove `include/exclude` directives and implicit bonds from the model	identification of control relationships through iterative processing
3. binding between shared and unique sites of the same auxiliary protein	complexes that contain a bond or a chain of bonds that connects the shared and unique sites of one auxiliary protein or two auxiliary proteins of the same type	generates reduced models with incorrect mass-balance	inspect species list for occurrence of such complexes. validate results of reduced model against full model	automated detection of offending complexes
4. control relationships between sites on different proteins	state of protein *P* influences transformation between states of protein *Q* within a complex	generation of incorrect reduced models for the observables that contain *Q*	disable reduction of *P* by using the command-line option '`-nored P`'	unknown
5. binding between shared sites of different proteins	binding of reducible proteins through shared sites, that is, heterodimerisation	generation of incorrect reduced models	disable reduction of proteins involved in heterodimerisation using `–nored` option	unknown
6. stochastic simulations	simulation of reduced models using kinetic Monte Carlo methods, for example, Gillespie algorithm	levels of shared site occupancy are decoupled among auxiliary proteins	validate results of reduced model against full model to estimate size of errors	correlated Monte Carlo sampling [33]

See corresponding section of Supplement 5 for further discussion of each limitation

sites. If such species are found, the user must manually disable macro-reduction of the involved protein using the `–nored` option (see more details in Section 3 of Supplement 5).

4. Control relationships between sites on different proteins: The algorithm presented here only utilises hierarchical control relationships within a single protein. However, for models in which the state of one protein in a complex affects the transformations between the states of another protein in the same complex, the use of only within-protein control relationships may give an incorrect reduced model. An example involving ordered phosphorylation of an adapter protein is illustrated in Fig. S5.2, and a `bngl`-file for the model is given in Supplement 6. The current implementation does not detect control relationships between sites of different proteins, and if such relationships exist in a model, the user is also advised to disable domain-oriented reduction (`–nored` option). In addition, incorrect model reduction can be detected by comparing time courses of the reduced and exact models, as shown in Fig. S5.3.

5. Binding between shared sites of different proteins: The algorithm does not identify reductions when auxiliary proteins from different progenitor proteins bind to each other through shared sites. This case is important because when multi-site signalling proteins can form dimers, formation of the dimer frequently modulates the activity of sites within each protein – a prominent example being the ErbB family of RTKs [30, 31]. Allowing association of the auxiliary proteins of one progenitor protein with the auxiliary proteins of a different progenitor leads to a proliferation in the number of heterodimers. The resulting problem of generating the correct mass balances for the binding partners of the progenitor proteins is not solved by the simple trick that works in the case of homodimersiation, in which the complexes containing

IET Syst. Biol., 2008, Vol. 2, No. 5, pp. 342–351
doi: 10.1049/iet-syb:20070081

different auxiliary proteins of the same progenitor protein are simply removed from the model. To avoid the possibility of errors the user is advised to use the `-nored` option for the proteins that undergo heterodimerisation.

6. Stochastic simulations: 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 [32] 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 a binding or dissociation event occurs involving a shared site, the event should apply to all of the shared sites of the same molecule. In the macro model each of these events will be governed by a separate reaction (albeit with the same rate) and these will fire independently in a stochastic simulation, which de-correlates the levels of shared site occupancy for the auxiliary proteins sharing the site. (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 [33], 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 (Appendix).

3 Discussion

The multiplicity of scaffold proteins involved in RTK signalling 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 [34, 35]. 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 signalling networks, a microscopic 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 signalling networks.

Based on these findings, we have developed a method for automatic domain-oriented reduction of signalling network models, which is implemented as a module in the software package BioNetGen. The reduction module takes a standard `bngl`-file as input and performs the following steps (Fig. 2). 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 in increased computational efficiency with no loss of accuracy (Table 1). The method is fully automated, and the reduction module takes as input a standard BioNetGen input file including standard simulation commands (Appendix and [19]). The only difference in output between a standard BioNetGen simulation and one run through the macro module is that species concentrations are reported only for the macro variables and not for the microscopic species. Time courses of observables generated by the full and reduced models will be identical, except in the cases noted in Section 2.4. Use of the module does not require the user to understand details of the algorithm, although the user is required to recognise the possible pitfalls described in Section 2.4 and in some cases to manually turn off reduction of problem proteins. Future work will focus on overcoming limitations to the applicability of the algorithm outlined in Section 2.4 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 Acknowledgments

The authors would like to thank Dr. Mikhail V. Kravchenko from the Burnasyan Federal Medical Biophysical Center in Moscow, Russia, who substantively rewrote, rearranged and optimised 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.

5 References

[1] BORISOV N.M., MARKEVICH N.I., HOEK J.B., KHOLODENKO B.N.: 'Signaling through receptors and scaffolds: independent interactions reduce combinatorial complexity', *Biophys. J.*, 2005, **89**, pp. 951–66

[2] BORISOV N.M., MARKEVICH N.I., HOEK J.B., KHOLODENKO B.N.: 'Trading the micro-world of combinatorial complexity for the macro-world of protein interaction domains', *BioSyst.*, 2006, **83**, pp. 152–66

[3] BURACK W.R., SHAW A.S.: 'Signal transduction: hanging on a scaffold', *Curr. Opin. Cell Biol.*, 2000, **12**, pp. 211–216

[4] CONZELMANN H., SAEZ-RODRIGUEZ J., SAUTER T., KHOLODENKO B.N., GILLES E.D.: 'A domain-oriented approach to the reduction of combinatorial complexity in signal transduction networks', *BMC Bioinformatics*, 2006, **7**, p. 34

[5] HLAVACEK W.S., FAEDER J.R., BLINOV M.L., PERELSON A.S., GOLDSTEIN B.: 'The complexity of complexes in signal transduction', *Biotechnol. Bioeng.*, 2003, **84**, pp. 783–794

[6] LEVCHENKO A., BRUCK J., STERNBERG P.W.: 'Scaffold proteins may biphasically affect the levels of mitogen-activated protein kinase signaling and reduce its threshold properties', *Proc. Natl. Acad. Sci. USA*, 2000, **97**, pp. 5818–5823

[7] SCHLESSINGER J.: 'Cell signaling by receptor tyrosine kinases', *Cell*, 2000, **103**, pp. 211–225

[8] DE MEYTS P., WHITTAKER J.: 'Structural biology of insulin and IGF1 receptors: implications for drug design', *Nat. Rev. Drug Discov.*, 2002, **1**, pp. 769–83

[9] PAWSON T., NASH P.: 'Assembly of cell regulatory systems through protein interaction domains', *Science*, 2003, **300**, pp. 445–452

[10] SCHLESSINGER J.: 'Common and distinct elements in cellular signaling via EGF and FGF receptors', *Science*, 2004, **306**, pp. 1506–1507

[11] TYSON J.J., NOVAK B., ODELL G.M., CHEN K., THRON C.D.: 'Chemical kinetic theory: understanding cell-cycle regulation', *Trends Biochem. Sci.*, 1996, **21**, pp. 89–96

[12] GOLDSTEIN B., FAEDER J.R., HLAVACEK W.S.: 'Mathematical and computational models of immune-receptor signalling', *Nat. Rev. Immunol.*, 2004, **4**, pp. 445–456

[13] FAEDER J.R., BLINOV M.L., GOLDSTEIN B., HLAVACEK W.S.: 'Combinatorial complexity and dynamical restriction of network flows in signal transduction', *Syst. Biol. (Stevenage)*, 2005, **2**, pp. 5–15

[14] FAEDER J.R., BLINOV M.L., GOLDSTEIN B., HLAVACEK W.S.: 'Rule-based modeling of biochemical networks', *Complexity*, 2005, **10**, pp. 22–41

[15] KOSCHORRECK M., CONZELMANN H., EBERT S., EDERER M., GILLES E.D.: 'Reduced modeling of signal transduction - a modular approach', *BMC Bioinformatics*, 2007, **8**, p. 336

[16] DANOS V., FERET J., FONTANA W., HARMER R., KRIVINE J.: 'Rule-based modelling of cellular signalling', *Lect. Notes Comput. Sci.*, 2007, **4703**, pp. 17–41

[17] KIYATKIN A., AKSAMITIENE E., MARKEVICH N.I., BORISOV N.M., HOEK J.B., KHOLODENKO B.N.: 'Scaffolding protein Grb2-associated binder 1 sustains epidermal growth factor-induced mitogenic and survival signaling by multiple positive feedback loops', *J. Biol. Chem.*, 2006, **281**, pp. 19925–19938

[18] BLINOV M.L., FAEDER J.R., GOLDSTEIN B., HLAVACEK W.S.: 'BioNetGen: software for rule-based modeling of signal transduction based on the interactions of molecular domains', *Bioinformatics*, 2004, **20**, pp. 3289–3291

[19] FAEDER J.R., BLINOV M.L., HLAVACEK W.S.: 'Rule-based modeling of biochemical systems with BioNetGen' in MALY I.V. (ED.): 'Methods in Molecular Biology: Systems Biology' (Humana Press, Totowa, NJ, 2008)

[20] LOK L., BRENT R.: 'Automatic generation of cellular reaction networks with Moleculizer 1.0', *Nat. Biotechnol.*, 2005, **23**, pp. 131–136

[21] MORTON-FIRTH C.J., BRAY D.: 'Predicting temporal fluctuations in an intracellular signalling pathway', *J. Theor. Biol.*, 1998, **192**, pp. 117–128

[22] MORTON-FIRTH C.J., SHIMIZU T.S., BRAY D.: 'A free-energy-based stochastic simulation of the Tar receptor complex', *J. Mol. Biol.*, 1999, **286**, pp. 1059–1074

[23] LE NOVERE N., SHIMIZU T.S.: 'STOCHSIM: modelling of stochastic biomolecular processes', *Bioinformatics*, 2001, **17**, pp. 575–576

[24] BLINOV M.L., YANG J., FAEDER J.R., HLAVACEK W.S.: 'Graph theory for rule-based modeling of biochemical networks', *Lec. Notes Comput. Sci.*, 2006, **4230**, pp. 89–106

doi: 10.1049/iet-syb:20070081

[25] DANOS V., LANEVE C.: 'Formal molecular biology', *Theor. Comput. Sci.*, 2004, **325**, pp. 69–110

[26] KHOLODENKO B.N., DEMIN O.V., MOEHREN G., HOEK J.B.: 'Quantification of short term signaling by the epidermal growth factor receptor', *J. Biol. Chem.*, 1999, **274**, pp. 30169–30181

[27] BLINOV M.L., FAEDER J.R., GOLDSTEIN B., HLAVACEK W.S.: 'A network model of early events in epidermal growth factor receptor signaling that accounts for combinatorial complexity', *BioSyst.*, 2006, **83**, pp. 136–151

[28] FAEDER J.R., HLAVACEK W.S., REISCHL I., *ET AL.*: 'Investigation of early events in FcεRI-mediated signaling using a detailed mathematical model', *J. Immunol.*, 2003, **170**, pp. 3769–3781

[29] BARUA D., FAEDER J.R., HAUGH J.M.: 'Structure-based Kinetic Models of Modular Signaling Protein Function: Focus on Shp2', *Biophys. J.*, 2007, **92**, pp. 2290–2300

[30] BIRTWISTLE M.R., HATAKEYAMA M., YUMOTO N., OGUNNAIKE B.A., HOEK J.B., KHOLODENKO B.N.: 'Ligand-dependent responses of the ErbB signaling network: experimental and modeling analyses', *Mol. Syst. Biol.*, 2007, **3**, p. 144

[31] YARDEN Y., SLIWKOWSKI M.X.: 'Untangling the ErbB signalling network', *Nat. Rev. Mol. Cell Biol.*, 2001, **2**, pp. 127–137

[32] GILLESPIE D.T.: 'A general method for numerically simulating the stochastic time evolution of coupled chemical reactions', *J. Comp. Phys.*, 1976, **22**, pp. 403–434

[33] YERMAKOV S.M., MIKHAILOV G.A.: 'Statistical Modelling' (Nauka, Moscow, 1982)

[34] DANOS V., FERET J., FONTANA W., KRIVINE J.: 'Scalable Simulation of Cellular Signalling Networks', *Lect. Notes Comput. Sci.*, 2007, **4807**, p. 139

[35] Yang J., Monine M.I., Faeder J.R., Hlavacek W.S.: 'Kinetic Monte Carlo method for rule-based modeling of biochemical networks', arXiv:0712.3773, 2007

[36] HUCKA M., FINNEY A., SAURO H.M., *ET AL.*: 'The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models', *Bioinformatics*, 2003, **19**, pp. 524–531

6 Appendix: overview of BioNetGen

BioNetGen provides a flexible language for the description of protein structure and protein interactions called the BioNetGen language (BNGL) [19]. A model specification in the `bngl`-file may consist of five required elements: `parameters`, `species` (also called `seed species`), `reaction rules`, `observables`, and `actions`. Parameters specify the kinetic rate constants, total protein concentrations and other fixed numerical properties of the model. Species describe molecules (including their sites and states of these sites) that are present at the start of network generation. For example, the species `P(s1,s2~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 in the state named pY (a mnemonic for phosphotyrosine) and also free.

Reaction rules list the rules for building the biochemical network. For example, the reversible rule

$$\texttt{A(b)} + \texttt{B(a, c)} < - > \texttt{A(b!1) . B(a!1, c) k\_on, k\_off}$$

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 $k_{\texttt{on}}$, and all dissociation reactions will have the first-order rate constant $k_{\texttt{off}}$. 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

$$\texttt{Molecules P\_s2\_phos P(s2} \sim \texttt{pY)}$$

defines the observable named `P_s2_phos` of type `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 last major element of a `bngl`-file is the set of actions, which are commands that operate on a model specification. Two basic commands are illustrated in the examples presented in Supplements 1, 2, and 6. The `generate_network` command automatically generates the set of 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 (`net`-file) or exported in the Systems Biology Markup Language [36], which can be imported by a large number of other simulation and analysis tools. The `simulate_ode` command performs an ODE-based simulation of the network over a specified time period with results reported at specified time points. Additional commands and details of BNGL syntax can be found in [19].