

RULE-BASED MODELING IN BIOUML

N. Mandrik^{*1,2,3}, *E. Kutumova*^{1,2}, *F. Kolpakov*^{1,2}

¹BIOSOFT.RU, LLC, Novosibirsk, Russia

²Design Technological Institute of Digital Techniques, SB RAS, Novosibirsk, Russia

³Sobolev Institute of Mathematics, SB RAS, Novosibirsk, Russia

e-mail: manikitos@biosoft.ru

*Corresponding author

Key words: rule-based modeling, BioUML, KaSim, BioNetGen

Abstract

Motivation and Aim: The traditional approach to mathematical modeling of biological systems involves usage of nonlinear systems of ordinary differential equations (ODEs) with given initial conditions. Talking about the modeling, we emphasize the fact that we consider an abstraction and study the mathematical description of some qualitative and quantitative characteristics of biological processes. The level of detail is dependent on the problem and based on the knowledge of the researcher. On the one hand, many meaningful models consist of few nonlinear equations. On the other hand, a detailed study of the biochemical networks leads to development of large-scale models consisting of hundreds of variables and, therefore, equations. Moreover, if we incorporate to the model site-specific details of protein-protein interactions, the number of protein modifications increases dramatically, and complexity of the model becomes combinatorial. For example, a protein comprising n amino acids can be potentially found in 2^n distinct phosphorylation states.

Investigation of such models using formalism of differential equations is difficult in view of the fact that we need to analyze thousands of variables whose values are often small. Visualization of the models (graphical representation of the reaction network as diagram) using one of the conventional standards (e.g., SBGN or KEGG) does not simplify the problem, although the diagram is easier to interpret than the corresponding system of equations, and readability of the diagram can be improved.

Methods and Algorithms: The main idea to deal with such models is based on representations of protein-protein interactions using rules serving as generators of species and biochemical reactions (or discrete events). This approach is known as «rule-based» modeling. Each rule describes a class of reactions with a common kinetic law and establishes the correspondence between reactant and product patterns defining a set of species with similar chemical compositions and properties.

Conclusion: The principles for creation of the «rule-based» models were implemented in several software resources including KaSim (<http://dev.executableknowledge.org/>) and

BioNetGen (www.bionetgen.org). BioUML supports the BioNetGen language (BNGL) and a special graphical notation created on the basis of SBGN and use it to visualize the «rule-based» models.

Availability: BioUML is available for download from www.biouml.org (free).

Introduction

The traditional approach to mathematical modeling of biological systems involves usage of nonlinear systems of ordinary differential equations (ODEs) with given initial conditions. Talking about the modeling, we emphasize the fact that we consider an abstraction and study the mathematical description of some qualitative and quantitative characteristics of biological processes. The level of detail is dependent on the problem and based on the knowledge of the researcher. On the one hand, many meaningful models consist of few non-linear equations. On the other hand, a detailed study of the biochemical networks leads to development of large-scale models consisting of hundreds of variables and, therefore, equations [1]. Moreover, if we incorporate to the model site-specific details of protein-protein interactions, the number of protein modifications increases dramatically, and complexity of the model becomes combinatorial. For example, a protein comprising n amino acids can be potentially found in 2^n distinct phosphorylation states [2].

The model by Kholodenko *et al.* [3] with extensions [4–6] shows more meaningful example. This model describes relationships between the binding of epidermal growth factor (EGF) to its receptor (EGFR) at the cell surface and the activation of kinase cascade Ras/Raf/MEK/ERK through a series of adapter (Grb-2, Shc) and effector (Sos) proteins. Authors of the model do not consider the possible phosphorylation of 9 individual tyrosines of EGFR. If we do that, we need to take into account $2^9 = 512$ different phosphorylation states of an individual receptor and $2^9 + 2^8 \cdot (2^9 - 1) = 131328$ combinations of phosphorylation states of receptors in a dimer [7, 8].

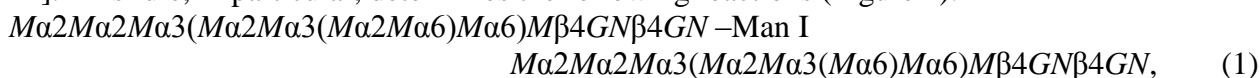
Krambeck *et al.* [9–11] give another example of a model with the combinatorial complexity. This model represents a reaction network of *N*-glycosylation process characterized by formation of *N*-glycans mediated by enzymatic removal or addition of monosaccharide (D-mannose, D-galactose, L-fucose, etc.) residues. The model begins with three high-mannose structures that lead to synthesis of about 10,000 – 20,000 glycans.

Investigation of such models using formalism of differential equations is difficult in view of the fact that we need to analyze thousands of variables whose values are often small. Visualization of the models (graphical representation of the reaction network as diagram) using one of the conventional standards (e.g., SBGN [12] or KEGG [13]) does not simplify the problem, although the diagram is easier to interpret than the corresponding system of equations, and readability of the diagram can be improved [14].

The main idea to deal with such models is based on representations of protein-protein interactions using rules serving as generators of species and biochemical reactions (or discrete events) [7]. This approach is known as «rule-based» modeling [8]. Each rule describes a class of reactions with a common kinetic law and establishes the correspondence between reactant and product patterns defining a set of species with similar chemical compositions and properties. For example, consider the rule of D-mannose removal from glycan structures mediated by Man I in the model by Krambeck *et al.* [10]:



Here, M denotes D-mannose, α_2 indicates the linkage between two of them, α is the linkage between D-mannose and some chemical structure defining a specific glycan (i.e. α_2 -, α_3 - or α_6 -linkage), and open parenthesis characterizes the beginning of a branch in the glycan structure [9–11]. This rule, in particular, determines the following reactions (Figure 1):



$\text{Ma}2\text{Ma}2\text{Ma}3(\text{Ma}2\text{Ma}3(\text{Ma}2\text{Ma}6)\text{Ma}6)\text{M}\beta4\text{GN}\beta4\text{GN} - \text{Man I}$
 $\text{Ma}2\text{Ma}3(\text{Ma}2\text{Ma}3(\text{Ma}2\text{Ma}6)\text{Ma}6)\text{M}\beta4\text{GN}\beta4\text{GN}, \quad (2)$

where the structures of glycans are given in accordance with the standard described by Banin *et al.* [15].

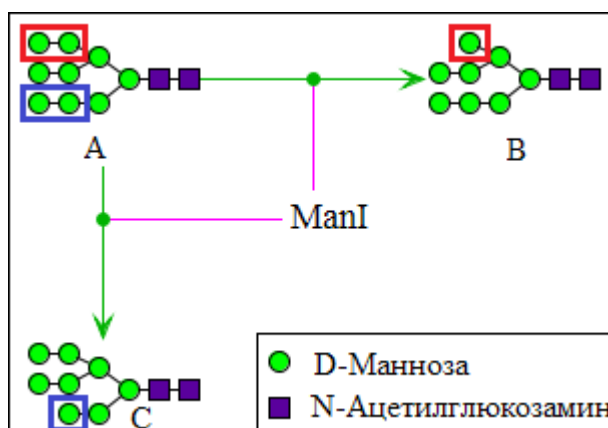


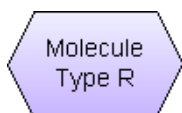
Figure 1. Visualization of biochemical reactions (1) and (2). Titles A, B, and C denote glycans $M2M2M3(M2M3(M2M6)M6)M4GN4GN$, $M2M2M3(M2M3(M6)M6)M4GN4GN$, and $M2M3(M2M3(M2M6)M6)M4GN4GN$, respectively. Green arrows represent reactions catalyzed by enzyme *Man I* (separated by magenta edges). Reactant patterns (in the structure of A) and product patterns (in the structures of B and C) for reactions (1) and (2) outlined by red and blue, respectively.

Knowing reaction rules and starting species of the model, we can get an ODE system for solution of which we can use the VODE solver applicable for stiff and nonstiff problems [16].

The principles for creation of the «rule-based» models were implemented in several software resources including KaSim (<http://dev.executableknowledge.org/>) and BioNetGen [2, 17]. To analyze the «rule-based» models in this work, we used the BioUML software (<http://biouml.org>) representing an open source Java-based integrated platform for visual modeling, formal description and analysis of complex biological systems. BioUML supports the BioNetGen language (BNGL) and a special graphical notation created on the basis of SBGN [12] and used to visualize the «rule-based» models (Table 1).

Table 1. Graphical notation for representation of «rule-based» models in BioUML.

Graphical notation	Description
	Molecule
	Molecule component
	Biochemical species or pattern (reactant or product) in a reaction rule (species graph)
	Reaction rule (square) and edges connecting it with reactant (the edge with arrow tip) and product patterns
	Observable parameter of the model



Molecule type using for the model validation based on the rules. If molecule types are defined, all molecules of the model are checked for the correspondence to one of them. If not, the check is not performed.

Creation of the «rule-based» models in BioUML

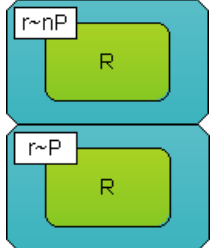
Description of the biochemical species

The special graphical notation related to BNGL can be used to define species of the «rule-based» models in BioUML (Table 2). Each species corresponds to a graph with nodes indicating chemical elements and edges denoting chemical bonds (Figure 2). Identical species correspond to isomorphic graphs. Therefore, we can avoid addition of such species to the model. Note that a reaction rule is applicable to a species, if the graph of its reactant pattern coincides with a subgraph of the species graph.

Table 2. Graphical representation of the BioNetGen formulas in BioUML.

BioNetGen formulas	Description	Graphical notation in BioUML
$R(r,r)$	Species (or pattern) named R with two binding sites r shown in parentheses after the name and separated by a comma.	
$R(r).L(l)^1$	Complex consisting of two species $R(r)$ and $L(l)$ indicates the existence of a chain of chemical bonds connecting them. Such notation is usually used in patterns of reaction rules or observable parameters.	
$R(r!1).L(l!1,l)$	The symbol “!” after the name of a binding site denotes the presence of a chemical bond. A unique identifier showing which sites are connected (number 1 in the formula) follows it.	
$R(r!+)^1$	The symbol “!+” after the name of a binding site also indicates the presence of a chemical bond. However, pattern does not contain information about specific species and binding site.	
$R(r!?)^1$	The symbol “!?” after the name of a binding site is applied when the presence of a chemical bond is not exactly known.	

¹The marked formulas are used only in patterns of reaction rules or observable parameters. Other formulas can be used in the reaction rules as well as for the species definition.

$R(r \sim nP)$, $R(r \sim P)$	<p>The symbol “~” separates the name and state (e.g. phosphorylation) of a binding site. In the example formulas, P and nP denote phosphorylated and not-phosphorylated states of the binding site r, respectively.</p>	
-----------------------------------	--	---

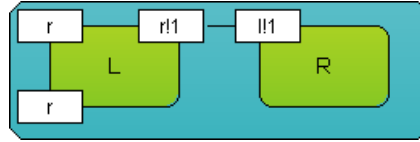
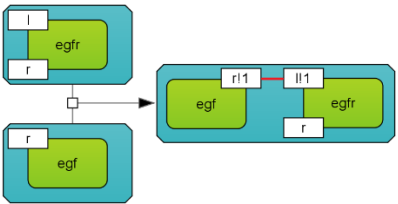
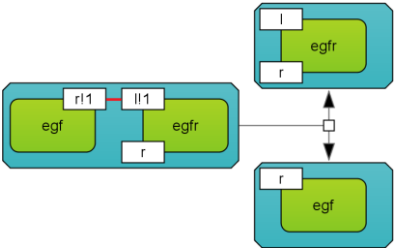
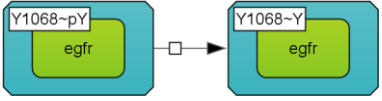


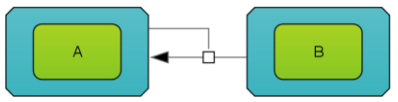
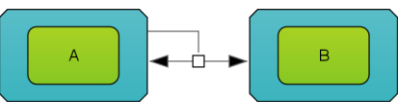
Figure 2. Graphical representation of the complex $L(r,r,r!1).R(!1)$.

Formalization of the rules

Each rule defines a transformation of graphs of reactant patterns. There are five major transformations [2, 17]: edge addition, edge removal, changing of a node state, node addition, and node removal (Table 3). Knowing transformation defined by a rule, we can identify a biochemical change of molecules induced by the corresponding reaction. For example, if a rule describes the addition of edge between two nodes of a graph (or graphs), then the reaction created using this rule corresponds to formation of new chemical bond (e.g. association of two molecules into a complex).

Table 3. The major transformations of the graphs determined by reactant patterns of reaction rules.

Graph transformation	Biochemical change of molecules	Example
Edge addition	Formation of new chemical bond	<p>Addition of the picked out edge</p> 
Edge removal	Disintegration of the chemical bond	<p>Removal of the picked out edge, dissociation of the complex</p> 
Changing of a node state	Changing of the binding site	<p>Dephosphorylation of EGFR</p> 
Node removal (perhaps, the whole graph)	Removal of a molecule	Reaction rule $A + B \rightarrow A$ describing the

		<p>removal of a molecule B.</p> 
Node addition (perhaps, in the new graph)	Addition of a molecule	<p>Reaction rule $A \rightarrow A + B$ describing the addition of a molecule B.</p> 

Transformation of a graph is directly defined by differences between reactant and product patterns in the rule. Analyzing these differences, we can find correspondence between molecules related to these patterns and can perform suitable transformation.

Creation of the complete reaction network

In addition to the set of rules in the rule-based models, we need to specify a set of starting species S_0 ("seed species" in the BioNetGen notation). Creation of the reaction network is an iterative process. In the first step, we apply each rule of the model to S_0 by the following way.

1. We get a set of reactants representing species from S_0 , which graphs include reactant pattern of the rule as subgraph.
2. Using the set of reactants and graph transformation defined by the rule for reactant patterns, we find a set of products.
3. Based on the possible constraints or inconsistencies between resulting products and product patterns in the rule, we eliminate a part of resulting reactions.

In the next step with the index $i \in \{1, 2, \dots\}$, we apply model rules to the set $S_i = S_{i-1} \cup P_{i-1}$, where P_{i-1} denotes the set of products found in the previous step.

The iterations are terminated when one of the following conditions is performed.

1. We achieve the maximal number of reactions or species.
2. We achieve the maximal number of iteration steps.
3. We did not find new reactions on the current step of iterations.

Parameters of the model

Two kind of parameters can be determined in the «rule-based» models:

1. kinetic parameters modulating rates of reactions induced by the model rules;
2. observable parameters representing sums of reactant concentrations corresponding to the specific reactant patterns; these parameters are used for analysis of the solution of the ODE system obtained for the complete reaction network.

Examples

Example 1

Consider a model consisting of two starting species $R(r,r)$ and $L(l,l)$, and three reaction rules:

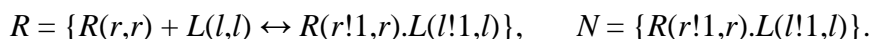
1. Formation of the complex: $R(r) + L(l,l) \leftrightarrow R(r!1).L(l!1,l)$.
2. Extension of the reaction chain: $R(r) + L(l,l!+) \leftrightarrow R(r!1).L(l!1,l!+)$.
3. Circuiting of the reaction chain: $R(r).L(l) \leftrightarrow R(r!1).L(l!1)$.

Graphs of species include edges connecting nodes of two different types (R and L). Each node has not more than two edges. In addition, we assume that we limited by two iterations by rules. Let introduce the following notations: S and N represent sets of all species and new species

generated in the i th iteration, respectively, R is a set of reactions, and RR_j^k is a set of reactants suitable for the reactant pattern k in the rule j .

First iteration. $S = \{R(r,r), L(l,l)\}$.

Using the first rule, we find $RR_1^1 = \{R(r,r)\}$ and $RR_1^2 = \{L(l,l)\}$ resulting in the new reaction:

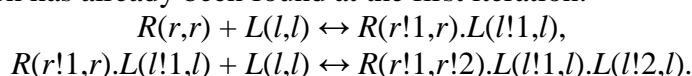


The symbol “!+” in the notation $L(l,!+)$ of the second rule means that the related node L of the reactant graph has one edge. Since $S = \{R(r,r), L(l,l)\}$, then we do not have suitable reactants, i.e. $RR_2^2 = \{\}$, although $RR_2^1 = \{R(r,r)\}$.

The similar situation is in the third rule. $R(r).L(l)$ indicates that the graph of a related reactant must include nodes R and L . Taking into account structure of the set S , we again do not have suitable reactants.

Second iteration. $S = \{R(r,r), L(l,l), R(r!1,r).L(l!1,l)\}$.

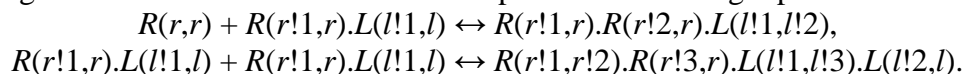
For the first rule, we derive $RR_1^1 = \{R(r,r), R(r!1,r).L(l!1,l)\}$, $RR_1^2 = \{L(l,l)\}$, and two reactions, one of which has already been found at the first iteration:



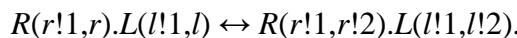
Therefore, we obtain

$$N = \{R(r!1,r!2).L(l!1,l).L(l!2,l)\}.$$

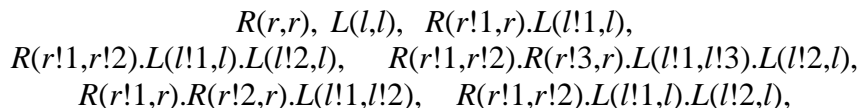
For the second rule, $RR_2^1 = \{R(r,r), R(r!1,r).L(l!1,l)\}$, $RR_2^2 = \{R(r!1,r).L(l!1,l)\}$. Thus, we get the following reactions and two elements for N represented in the right part of them:



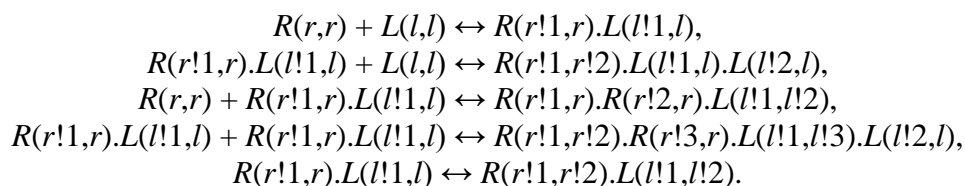
For the third rule, we obtain $RR_3^1 = \{R(r!1,r).L(l!1,l)\}$ inducing one reaction and refilling N by one new element:



The maximal number of iterations is achieved. As a result, we obtain the model consisting of seven species:



and five reactions:



Example 2

Consider the rule-based model of N -glycosylation created by Krambeck *et al.* [9–11]. This model consists of only three starting species (Figure 3) and 23 rules (Tables 4, 5) including possible constraints on structure of species (N -glycans) [8]. In addition to these constraints, authors specified adjustment coefficients used in kinetic laws and dependent on structure of reactants. Generation of the complete reaction network using these rules gives about 10,000–20,000 various N -glycans and about 30,000 reactions. We recreated the model using BioNetGen syntax and specified about 80 rules relating to the same reaction network.

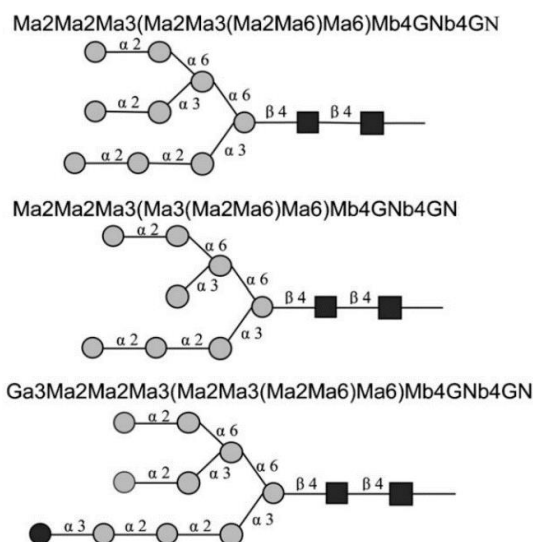


Figure 3. Starting species of the *N*-glycosylation model.

The relative difference between mass spectra calculated for the model by Krambeck *et al.* and the rule-based model created using BioNetGen syntax in the BioUML software is not more than 10%. The absolute difference is about 0.2% by total mass of all *N*-glycans in the model.

Table 4. Reaction rules of the *N*-glycosylation model.

Enzyme	Reactant	Product	Constraint
ManI	(Ma2Ma	(Ma	~*Ma3(...Ma6)Ma6 & ~Ga3
ManI	(Ma3(Ma2Ma3(Ma6)Ma6)	(Ma3(Ma3(Ma6)Ma6)	~Ga3
ManII	(Ma3(Ma6)Ma6	(Ma6Ma6	(GNb2 Ma3 & ~Gnbis
ManII	(Ma6Ma6	(Ma6	(GNb2 Ma3 & ~Gnbis
a6FucT	GNb4GN	GNb4(Fa6)GN	GNb2 Ma3 & #A=0 & ~Gnbis
GnTI	(Ma3(Ma3(Ma6)Ma6)Mb4	(GNb2Ma3(Ma3(Ma6)Ma6)Mb4	
GnTII	(GNb2 Ma3(Ma6)Mb4	(GNb2 Ma3(GNb2Ma6)Mb4	
GnTIII	GNb2 Ma3	GNb2 Ma3(GNb4)	~Ab & ~Gnbis
GnTIV	(GNb2Ma3	(GNb2(GNb4)Ma3	~Gnbis
GnTV	(GNb2Ma6	(GNb2(GNb6)Ma6	~Gnbis
iGnT	(Ab4GN	(GNb3Ab4GN	~*_Ma3 Mb4
b4GalT	(GN	(Ab4GN	~*GNb4(...Ma6)Mb4
a3SiaT	(Ab4GN	(NNA3Ab4GN	
IGnT	(Ab4GNb3Ab	(Ab4GNb3(GNb6)Ab	
a6SiaT	(Ab4GN	(NNA6Ab4GN	
b3GalT	(GN	(Ab3GN	~*GNb4(...Ma6)Mb4
FucTLe	Ab3GNb	Ab3(Fa4)GNb	
FucTLe	(...Ab4GNb	(Fa3(...Ab4)GNb	
FucTH	(Ab3GNb	(Fa2Ab3GNb	
FucTH	(Ab4GNb	(Fa2Ab4GNb	
a3FucT	(...Ab4GNb	(Fa3(...Ab4)GNb	
GalNAcT-A	(Fa2Ab	(Fa2(ANa3)Ab	
GalT-B	(Fa2Ab	(Fa2(Aa3)Ab	

Table 5. Codes for the reaction rules listed in the table 4.

Symbol	Meaning	Expression
...	Ligand	Any string (possibly empty) with all parentheses matched.
_	Continuation	Any string (possibly empty) where every "(" is matched with a following ")"
	Possible branch point	Empty string or "(...)"
*	Reaction site	Position of first difference between product and reactant string
Gnbis	Bisecting <i>GN</i>	Existence of the following substring Ma3(GNb4)(...Ma6)Mb4
#	Number of	
~	Logical not	

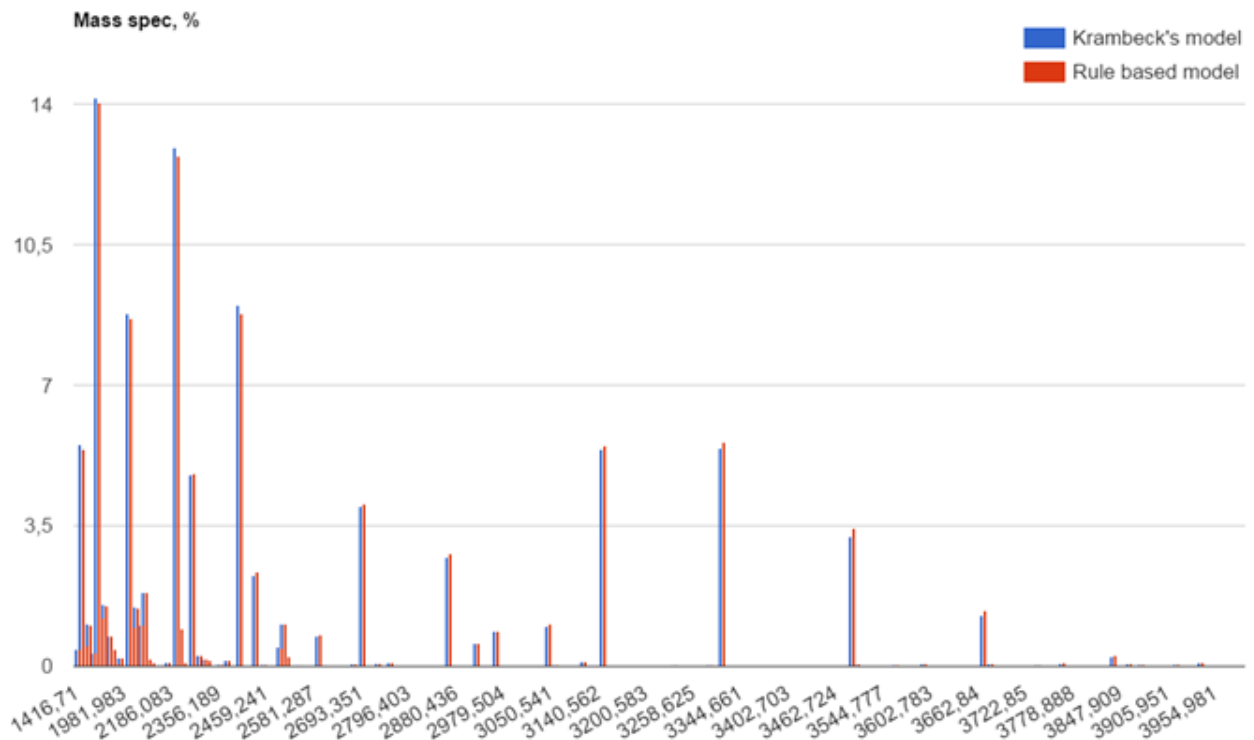


Figure 4. Comparison of mass spectra calculated for the model by Krambeck *et al.* and the rule-based model created using the BioUML software.

References

1. Hlavacek W.S. (2009) How to deal with large models? *Molecular Systems Biology*, 5:240.
2. Faeder J., Blinov M. and Hlavacek W. (2009) Rule-based modeling of biochemical systems with BioNetGen. *Methods Mol. Biol.* 500: 113-167.
3. Kholodenko B. N., Demin O. V., Moehren G., and Hoek J. B. (1999) Quantification of short term signaling by the epidermal growth factor receptor. *J. Biol. Chem.* 274, 30169–30181.
4. Schoeberl B., Eichler-Jonsson C., Gilles E. D. and Muller G. (2002) Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. *Nat. Biotechnol.* 20, 370–375.
5. Resat H., Ewald J. A., Dixon D.A., Wiley H. S. (2003) An integrated model of epidermal growth factor receptor trafficking and signal transduction. *Biophys J.* 85(2): 730-43.
6. Maly I.V., Wiley H.S., Lauffenburger D.A. (2004 Jan) Self-organization of polarized cell signaling via autocrine circuits: computational model analysis. *Biophys J.* 86 (1 Pt 1): 10-22.
7. Hlavacek, W. S., Faeder, J. R., Blinov, M. L., Posner, R. G., Hucka, M., and Fontana, W. (2006) Rules for modeling signal-transduction systems. *Sci. STKE* 2006, re6.
8. Blinov M.L., Moraru I.I. Leveraging Modeling Approaches: Reaction Networks and Rules. *Adv Exp Med Biol.* 2012 ; 736: 517–530.
9. Krambeck F.J., Betenbaugh M.J. A mathematical model of N-linked Biotechnology and Bioengineering. 2005. V. 92. № 6. P. 711–728.
10. Krambeck F.J., Bennun S.V., Narang S., Choi S., Yarema K.J., Betenbaugh M.J. A mathematical model to derive N-glycan structures and cellular enzyme activities from mass spectrometric data. *Glycobiology.* 2009. V. 19. № 11. P. 1163–1175.
11. Bennun S.V., Yarema K.J., Betenbaugh M.J., Krambeck F.J. Integration of the transcriptome and glycome for identification of glycan cell signatures. *PLoS Computational Biology.* 2013. V. 9. № 1. e1002813.
12. Le Novère N., Hucka M., Mi H., Moodie S., Schreiber F., Sorokin A., Demir E., Wegner K., Aladjem M.I., Wimalaratne S.M., et al. The Systems Biology Graphical Notation // *Nature Biotechnology.* 2009. V. 27, № 8. PP. 735-741.

13. Kanehisa M., Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, 2000, 28(1): 27–30.
14. Kitano H, Funahashi A, Matsuoka Y, Oda K. (2005) Using process diagrams for the graphical representation of biological networks. *Nat Biotechnol.* 23(8): 961-6.
15. Banin E., Neuberger Y., Altshuler Y., Halevi A., Inbar O., Nir D., Dukler A. A novel linear code (R) nomenclature for complex carbohydrates. *Trends in Glycoscience and Glycotechnology.* 2002. V. 14. № 77. P. 127–137.
16. Brown P.N., Byrne G.D., Hindmarsh A.C. VODE: A Variable-Coefficient ODE Solver // *SIAM Journal on Scientific and Statistical Computing.* 1989. V. 10. PP. 1038-1051.
17. Blinov M. L., Faeder J. R., Goldstein B. and Hlavacek W. S. (2004) BioNetGen: software for rule-based modeling of signal transduction based on the interactions of molecular domains. *Bioinformatics* 20: 3289-91.