A graphical notation for biochemical networks

Hiroaki Kitano

A solid definition and comprehensive graphical representation of biological networks is essential for efficient and accurate dissemination of information on biological models. Several proposals have already been made toward this aim. The most well known representation of this kind is a molecular interaction map, or 'Kohn Map'. However, although the molecular interaction map is a well-defined and compact notation, there are several drawbacks, such as difficulties in intuitive understanding of temporal changes of reactions and additional complexities arising from particular graphical representations. This article proposes several improvements to the molecular interaction map, as well as the use of the 'process diagram' to help understand temporal sequences of reactions.

Hiroaki Kitano
ERATO Kitano Symbiotic
Systems Project
JST and The Systems
Biology Institute
Suite 6A, M31, 6-31-15 Jingumae
Shibuya, Tokyo 150-0001, Japan
Sony Computer Science
Laboratories, Inc.
3-14-15 Higashi-Gotanda
Shinagawa, Tokyo 141-0022, Japan
e-mail: kitano@symbio.jst.go.jp

▼ Rapid progress in molecular biology, particularly high-throughput genomics and proteomics, continue to produce massive biological data. Extensive investigations on the molecules involved and their relationships are now being revealed, and researchers are starting to understand the network of genes and proteins. However, currently knowledge on molecular interactions is mostly described either by written text or by traditional cartoon-like diagrams. Written text is inherently ambiguous, and results have had to be re-interpreted by each reader of the article. Most authors of biological papers use arrow-headed lines and bar-headed lines to indicate activation and inhibition, respectively, with mixed and often inconsistent semantics. However, traditional diagrams are informal, often confusing, and much information is lost. Thus, the urgent task is to provide a set of notations that have powerful expression capability and are highly readable for biochemical and gene regulatory networks. Although there is a broad range of biological processes, the first step has to be made somewhere. The level of a biochemical network is a reasonable starting point because it is in the middle layer between elementary processes and organisms.

Traditional diagrams are ambiguous and ill-defined

In traditional schematics, symbols, such as arrows, are used in various semantics from transcriptional activation/inhibition to signal transduction via protein-protein interaction and transport, among others. This resulted in substantial confusion as researchers had to read substantial parts of the text to understand what the diagram really describes. In addition, it is often the case that both state transitions and relationships are described within one diagram, which makes the meaning of nodes and arcs confusing. One arrow may mean activation, but the other arrow in the same diagram may mean transition of the state or translocation of materials. Without consistent and unambiguous rules for representation, not only is information lost, but also the wrong information could be disseminated.

The situation is further aggravated by the lack of a standard for graphical notations and so schematics follow different rules preventing the efficient and accurate dissemination of knowledge. Recently, some researchers proposed notation systems to mitigate such problems [1-5], and many research groups working on large-scale biology have invented their own, but rather informal, system of describing networks and processes. Nevertheless, no standard has yet been formed. This is partly because proposed notations still have a number of issues to overcome, and there is also a lack of software tools to assist the use of such diagrams.

With abundant information on gene regulation, metabolic pathways and signal transduction, and increasing interest on system-level understanding of biological systems, there is a pressing need for powerful and standardized graphical notations to describe biological networks and processes. The absence of such representation potentially hinders systems-oriented

1478-5282/03/$ – see front matter ©2003 Elsevier Science Ltd. All rights reserved. PII: S1478-5282(03)02380-3

research because basic system-level information such as interactions in large-scale biochemical networks and gene regulations are not readily available and each research group has to make redundant efforts to reconstruct such information. It is evident and highly important that such representation is to be established and a standard is formed. At the same time, this representation cannot be developed overnight as it has to be able to represent a wide range of biological processes and accommodate the diverse needs of researchers. A system of graphical representation should be powerful enough to express sufficient information in a clearly visible and unambiguous way and should be supported by software tools.

Therefore, the aim of this paper is to briefly analyze issues with current notation and to propose a set of ideas and improvements, thereby contributing to the progress of the field and eventual standard formation.

Requirements for graphical notation

There are several criteria that the notation system should satisfy, such as:

(1) **Expressiveness:** The notation system should be able to describe every possible relationship among genes and proteins, as well as biological processes as a whole.

(2) **Semantically unambiguous:** Notation should be unambiguous. Different semantics should be assigned to different symbols that are clearly distinguishable.

(3) **Visually unambiguous:** Each symbol should be clearly identified and not be mistaken with other symbols. This feature should be maintained with low-resolution displays, using only black and white.

(4) **Extension capability:** The notation system should be flexible enough to add new symbols and relationships in a consistent manner. This may include the use of color-coding to enhance expressiveness and readability, but information should not be lost even with black and white displays.

(5) **Mathematical translation:** The notation should be able to convert itself into mathematical formalisms, such as differential equations, so that it can be directly applied for numerical analysis.

(6) **Software support:** The notation should be supported by software for its drawing, viewing, editing and translation into mathematical formalisms.

Molecular interaction map (MIM) is the most rigidly defined diagram to date that describes interactions among molecular species [1,2] that meets many, but not all, of the above requirements. Each protein is represented as a node and appears only once in the diagram. Various relationships are defined with different symbols that can be distinguished from each other. This notation is solid and well designed, and most importantly, a large network of the mammalian cell cycle has already been described using MIM. This notation provides a rigid framework to describe relationships among species, and with a series of modifications and extensions it can be the basis for standard representation. Some drawbacks of MIM, however, include (1) difficulties in understanding the process of certain reactions because the temporal order of reactions are represented implicitly, (2) complexity of the diagram because effects of residue modifications are represented as connecting lines outside of the node, (3) lack of notation for complex cis-regulations, and (4) the need for sufficient experiences to correctly draw and read compound formation diagrams, partly caused by the absence of software tools to support this diagram drawing. In the rest of the paper, a set of improvements based on MIM will be presented, followed by an introduction into the process diagram that complements MIM by explicitly representing temporal sequences of biological processes.

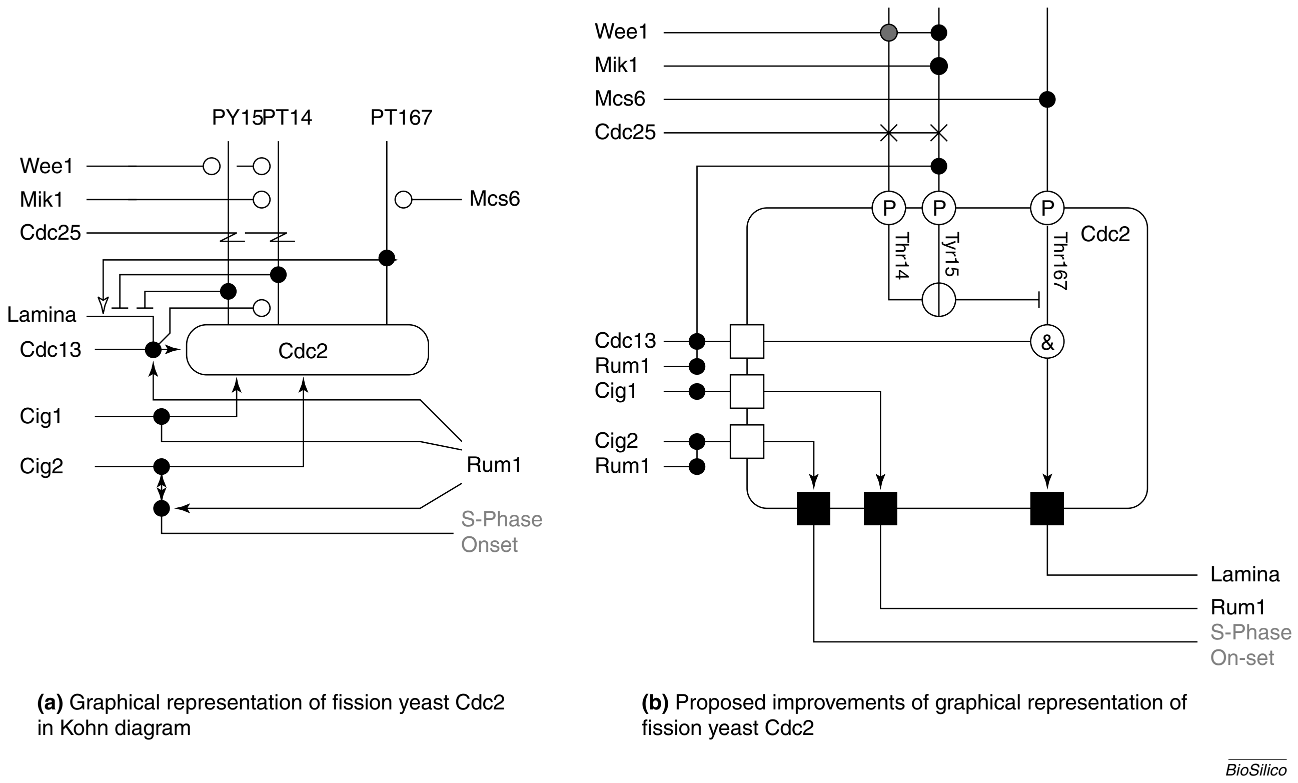
Modifying the molecular interaction map

One of the difficulties in understanding MIM comes from the way in which effects of residue modifications and protein bindings are described. In MIM, such effects are described using arrow-headed and bar-headed lines surrounding the node. This causes crowded graphics and extensive line crossings that make the diagram less understandable. Cook and coworkers suggested that such effects are represented inside of the node [3] so that graphics are less crowded. Cook's notation, however, does not represent specific residue and binding-site positions, which is essential information for understanding molecular interactions. This notation can be incorporated into MIM.

Figure 1 shows a graphical representation of interactions involving the fission yeast protein Cdc2 in (a) the original MIM and (b) a modified diagram showing improvements. In the modified diagram, the way in which residue modifications and protein binding change kinase activity of the protein are represented inside of the round-cornered box; all interactions regarding modification of residues are placed above the box and all binding relationships are described at the left and right side of the box

Interactions involving kinase activities and other effects are described below the box. Processes such as degradation and transport are reserved for the right side and below the box. The modified version of the diagram introduces a set of symbols so that modification, binding, internal logics and other changes in protein states can be described more clearly than in the original MIM (Fig. 2).

Figure 1 can be used as a building block for a diagram representing a network of interactions. Figure 3 is an



(a) Graphical representation of fission yeast Cdc2 in Kohn diagram

(b) Proposed improvements of graphical representation of fission yeast Cdc2

Figure 1. Representation of fission yeast Cdc2 protein in **(a)** the original MIM and **(b)** proposed improvements. Both diagrams represent interactions involving fission yeast Cdc2. Wee1 phosphorylates Thr14 and Tyr15, Mik1 phosphorylates Tyr15, Mcs6 phosphorylates Thr167, and Cdc25 dephosphorylates Thr14 and Tyr15. Cdc2 binds to either Cdc13, Cig1, or Cig2. When Cdc2 is forming a complex with Cdc13 and only Thr167 is phosphorylated, the complex interacts with Lamina. Phosphorylation of either Thr14 or Tyr15 inhibits activation of Cdc2 due to phosphorylation of Thr167. The complex auto-phosphorylates Try15 of its Cdc2. The complex of Cdc2 and Cig1 interacts with Rum1. Cdc2-Cdc13 complex and Cdc-Cig2 complex form heterotrimers involving Rum1.

example of such a diagram using the modified representation proposed. It should be noted that modular diagrams for proteins, each of which is distinguished by a dashed line box, are connected to form a network.

Extended versions of MIM can be used to describe logics behind cis-regulatory regions [6,7] as shown for the Endo-16 gene (Fig. 4). There are substantial internal logics in cis-regulation. This logic is a hybrid of boolian logic and

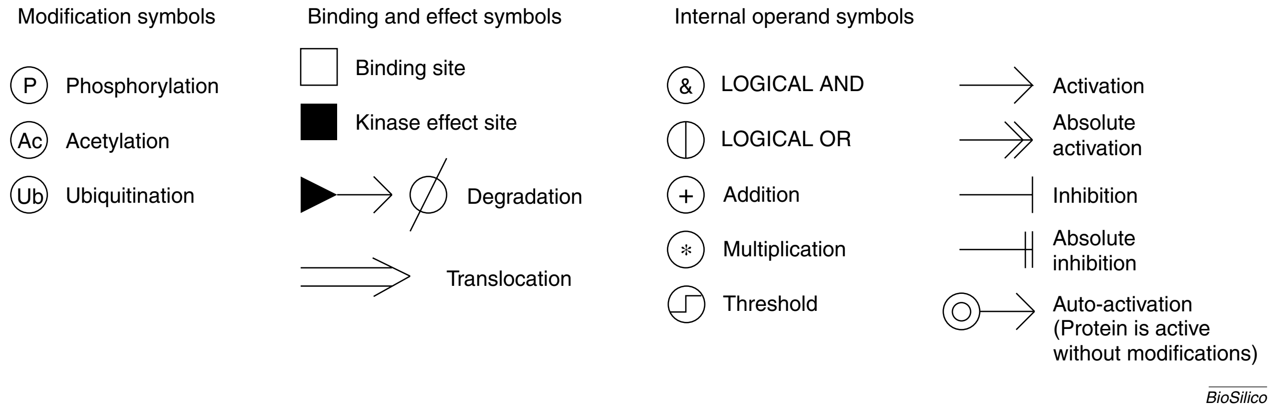
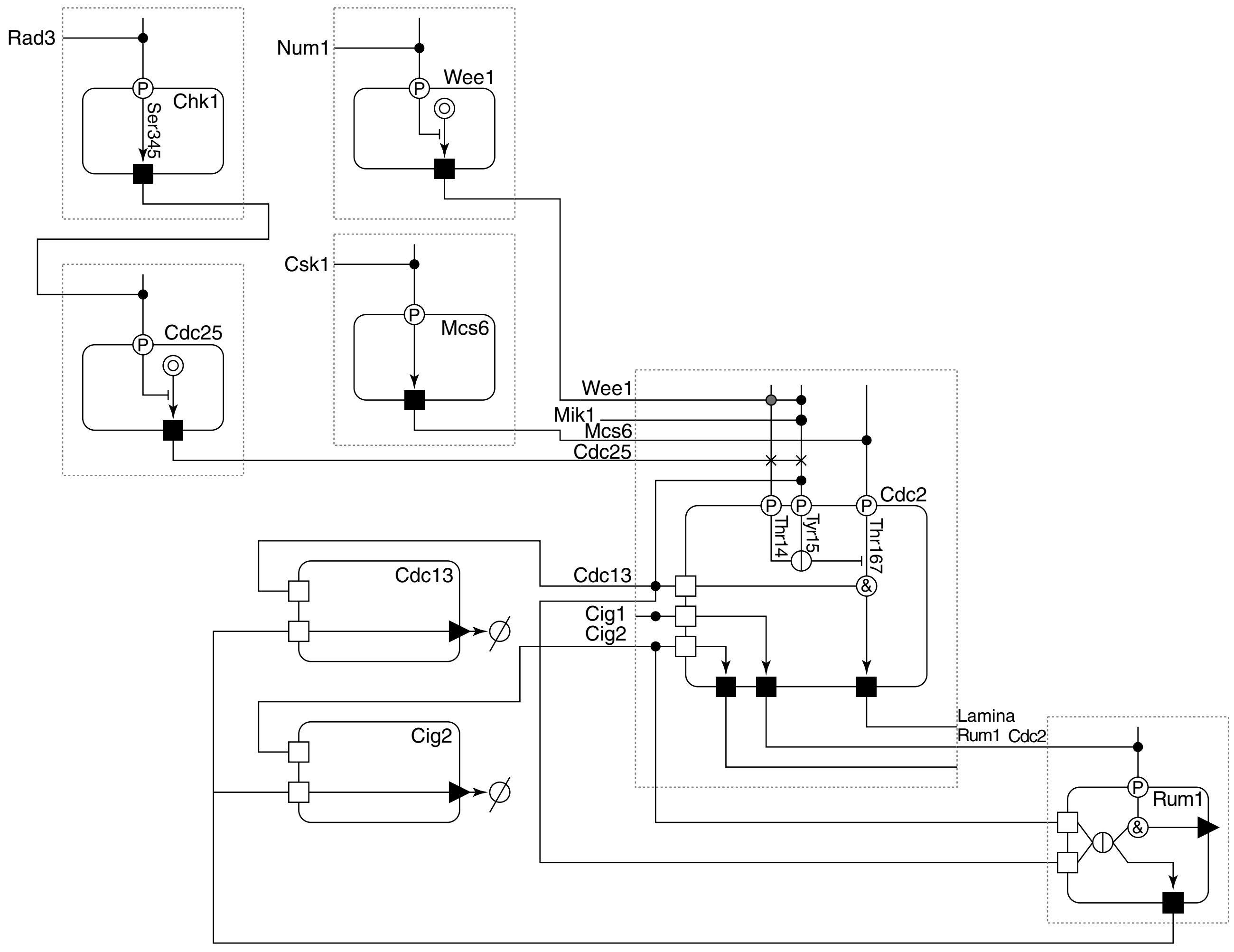
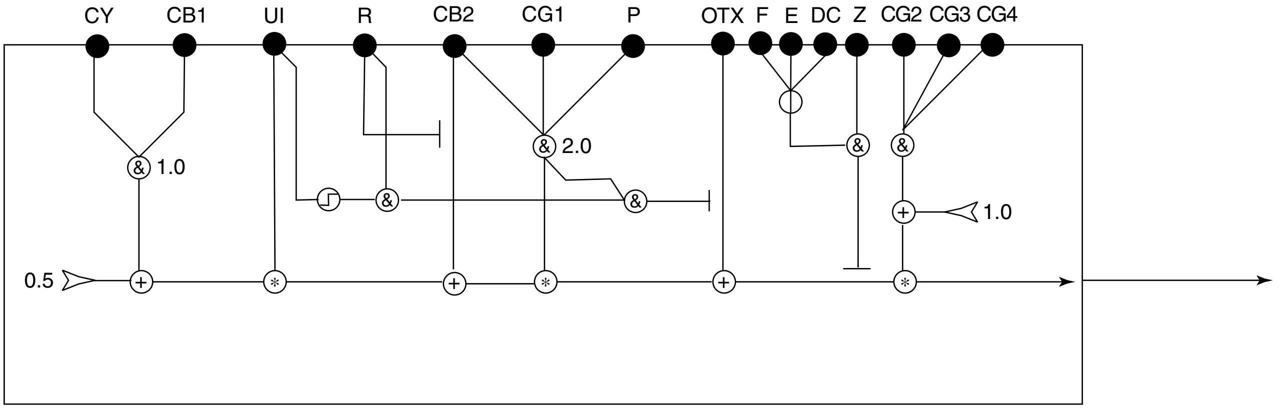


Figure 2. Symbols introduced in the modified version of the diagram. Absolute activation and absolute inhibition means that the protein is active or inactive, respectively, regardless of the other states. Auto-activation means the protein exhibits kinase activity when there is no modification. Without this symbol, the activity of protein without any modifications could be ambiguous.



BioSilico

Figure 3. Example of improved MIM describing a part of the fission yeast cell cycle. Interactions between Cdc2, Cdc13, Rum1, Mcs6, Wee1, Chk1 are described.



BioSilico

Figure 4. Example of representation for cis-regulation. Cis-regulation for the Endo-16 gene is represented as a possible extension of the Kohn diagram notation. Numbers associated with boolian logic operand symbols are output values when the condition of the logical operand is satisfied. Numbers associated with open triangles are bias values that should be added in the numeric addition operands.

numeric calculation, so that both logical and quantitative changes of transcriptional regulations are represented. In some nodes, numbers are associated to indicate output value of the node when logical value of the node is true. For example, output of the node, which is logical AND with inputs from CY and CB1, is 1.0 when both the CY and CB1 site is bounded by transcription factors.

Using this modified representation, allosteric reactions as seen in aspartate transcarbamoylase, can be represented by using internal operands among residues and binding sites. Figure 5 represents simple cases of allosteric reactions. Figure 5c shows part of an allosteric reaction in aspartate transcarbamoylase, which is one of the most well investigated systems with allostery and is literally a textbook example [8,9]. Several new notations are introduced. Internal protein binding sites are located inside the box and changes in these binding sites contribute to conformational change. Changes in conformation that are a mechanism behind allosteric reactions are not explicitly represented in the current scheme. Only the effects on binding sites and changes in kinase and other activities are represented.

Metabolic pathways constitute a large percentage of interacting networks of biological systems. Unlike signal transduction and the cell cycle, which have been used for examples so far, metabolic pathways are predominantly composed of reactions with mass action. Thus, a set of different symbols has to be introduced to denote this fact. In the extended MIM, special arrow-headed lines are used as seen in the reaction catalyzed by ATCase in Figure 5c. In general, the notation for metabolic pathways is relatively well agreed and consistent. Thus, a set of symbols for metabolic reactions share many commonly used notations in biochemistry text books [9,10]. There are also studies on notations for metabolic pathways [11]. Owing to space limitations, full explanations of notation for metabolic networks have to be left for other papers.

Maimon and Browning proposed diagrammatic cell language (DCL) with a set of improvements over MIM and different graphical symbols [5]. Notable features include

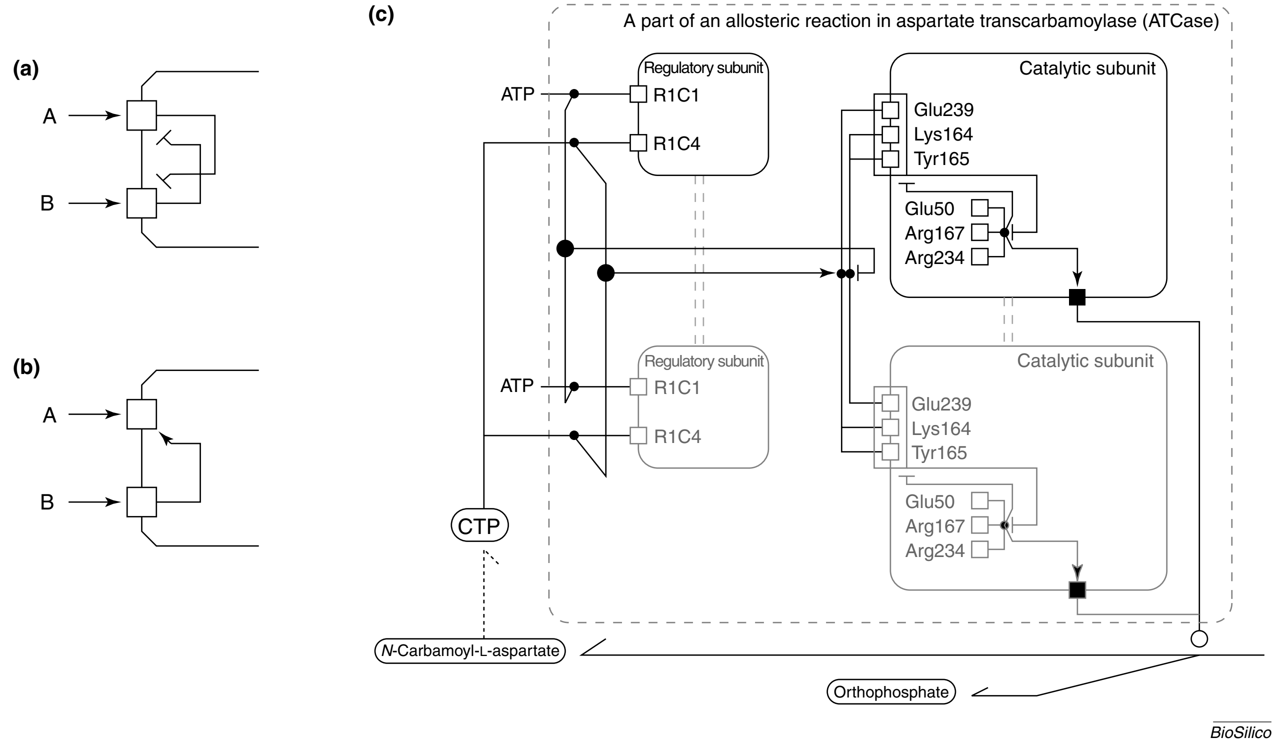


Figure 5. Representation of simple allosteric reactions. **(a)** In an inhibitory allosteric reaction, binding of either A or B causes conformational change that inhibits binding of other molecules. **(b)** In cooperative allosteric reactions, binding of B promotes binding of A. Changes in conformation are not explicitly represented. Only the consequences of such changes are represented as changes in binding capability. **(c)** Part of an allosteric interaction in aspartate transcarbamoylase is represented. For each catalytic subunit there are three binding sites (Glu239, Lys164 and Tyr165) which symmetrically binds to the other catalytic subunit under T state (inactive). Three internal binding sites (Glu50, Arg167 and Arg234) bind to each other causing a conformation change in R state (active). Formation of symmetrical binding of two catalytic subunits is promoted with CTP binding to regulatory subunits, and inhibited by ATP binding to the regulatory subunits. Gray-scaled boxes (a regulatory subunit and a catalytic subunit in low part of the diagram indicates that these are identical molecular species of the one defined above, connected by gray-scaled dashed parallel lines.

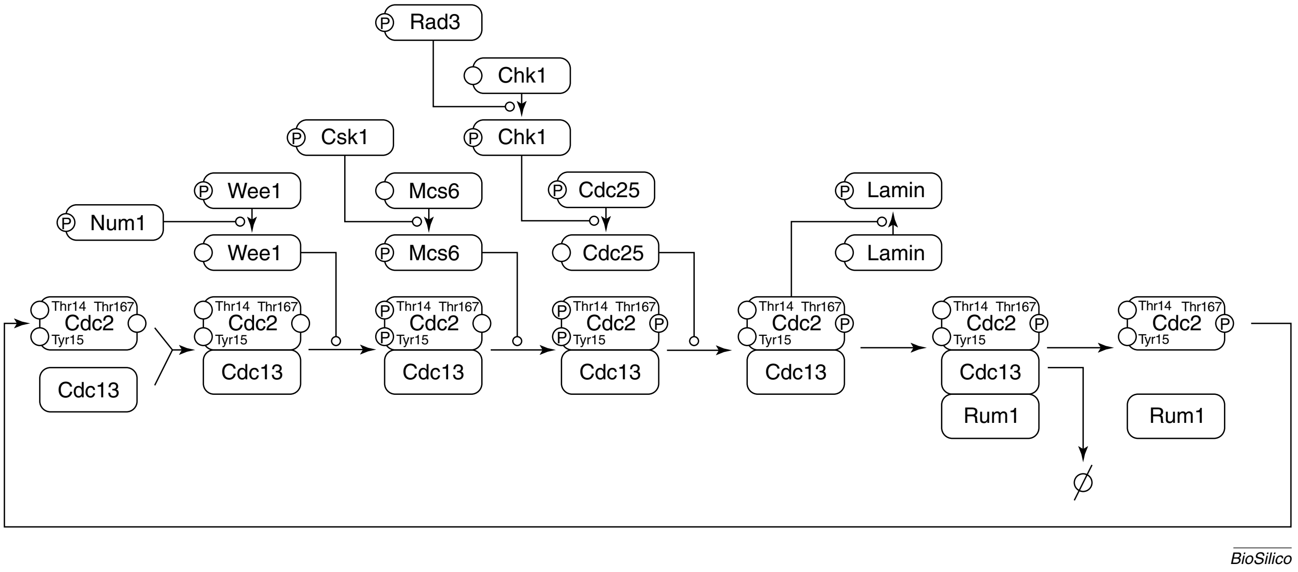


Figure 6. An example of the process diagram for part of the fission yeast cell cycle process represented in Figure 3. Temporal sequence of biochemical processes are represented explicitly. Molecular species appear repeatedly along the interaction processes.

the introduction of a link-box and a like-box. A link-box aims to represent binding sites and their status in a compact manner. A like-box is supposed to simplify the visualization of interactions that are similar to each other. In the proposed extension discussed in this paper, binding sites are explicitly represented on the boundary of the round-corner box that represents one molecular species. A like-box essentially describes the logical OR relationship for binding and reactions, which involves previously introduced internal logical symbols (Fig. 2).

These modifications proposed significantly enhance the ability of MIM to represent biological information and significantly improves the ease of understanding drawings graphical representations. However, even with the addition of proposed modifications, the notation is by no means complete. This is only a first step in the continuing process of improvements.

The process diagram

With the series of improvements and additions described above MIM is a good basis for a standard that rigidly represents interactions between molecular species, however, it does not explicitly show temporal sequences of biological events.

Trying to create a single diagram that satisfies all such needs is neither feasible nor desirable. An example of highly complex and dense networks is the 'very large scale integrated circuits' (VLSI) used in computer chips. In the VLSI design process, several diagrams with different levels of abstraction and emphasis are used to cover all necessary aspects of design details in a way that other readers can comprehend. A typical basic set of diagrams are the state-transition diagram, block diagram, flow chart, and timing chart with possible additions of data-flow and algorithm logic design diagrams. These diagrams represent different aspects of the same object. The MIM that defines the interaction between molecules within the system is very similar to the block diagram used in VLSI design, which implies that the addition of other diagrams may help in understanding complex biological systems.

The process diagram is used to visualize specific series of reactions among all possible reactions defined in the MIM. This helps researchers to intuitively understand specific temporal sequences of reactions that are more similar to informal cartoon-like diagrams commonly used today. Figure 6 shows an example of the process diagram that represents a part of the fission yeast cell cycle.

Apart from the obvious difference that this diagram explicitly represents a temporal sequence of reactions, there are several salient differences. First, molecular species appear repeatedly in the process. This is necessary because the process diagram explicitly represents the sequence of processes and how each molecular species changes during the process. Second, the process diagram does not have internal logic operand symbols in boxes representing molecular species. All logical operations are implicitly represented in the transition from one state to the other state during the reaction sequence.

There are several new symbols and representation that have to be introduced for the process diagram. First, a simplified representation of a complex is introduced. Unlike

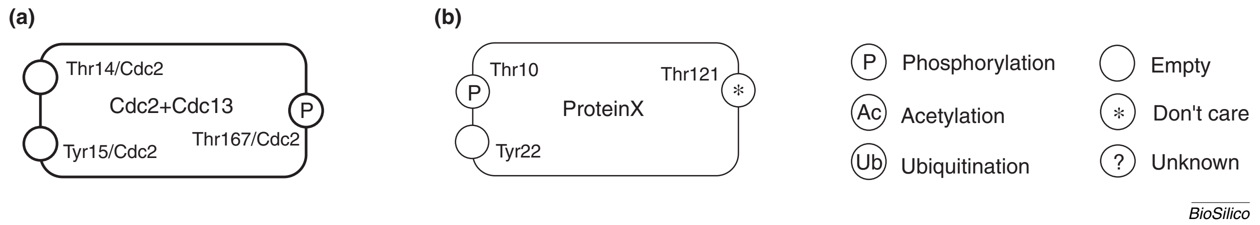


Figure 7. Protein representations in the process diagram. **(a)** A simplified complex of Cdc2 and Cdc13. **(b)** A hypothetical example of protein state. Thr10 is phosphorylated, Tyr22 is not modified and Thr121 is "don't care" meaning its state does not affect transition of the state of this protein.

the Kohn diagram, the process diagram requires molecular species and their complexes to appear repeatedly for every single reaction process that is represented. When there is a large complex of molecules involved, this feature soon becomes an obstacle for efficient and easy understanding, because large complexes dominate the visual representation and it becomes difficult to see the small changes. The countermeasure is to use simplified packing of complex representation. In Figure 6a, Cdc2 and Cdc13 that form a complex is represented as one box 'Cdc2+Cdc13'. Accordingly, residues are noted as 'position/protein', such as 'Thr14/Cdc2', so that whichever component of the complex residue is located it can be unambiguously identified. Such simplified representation significantly improves readability of the diagram.

Round symbols on the boundary of the box indicate the state of residues of interest. When each state and the activation state of the protein are known, all modifications on residues can be identified. However, it is often the case that when the states of residues affecting activation state are unclear, the symbol 'Unknown' is used. When the state of the residue does not affect activation state, the 'Don't Care' symbol is used to indicate this fact (Fig. 7b). Accordingly, three new residue symbols (empty, don't care and unknown) are introduced (Fig. 7c).

The process diagram is then consistent with the MIM, and in many cases it will represent a subset of interactions of the MIM. It should be possible to develop software that automatically generates the process diagram by specifying the starting state and the focus of reactions.

There are several mathematically well-established formalisms for describing parallel and concurrent processes. Petri net [12] is a typical example. Already there are a few studies using Petri net to describe and compute biological processes [13-15]. Once an extended MIM is translated into the process diagram, it should be straightforward to convert it into Petri net formalism for computation. Because the process diagram is intended to represent temporal sequences of biological processes, it does not define features like 'tokens' appearing in Petri net. Tokens are used for computing on the Petri net formalism, thus it is not necessary that MIM does not define a computing framework. Petri net has its own graphical representation of node and arc. Although it is a beautiful canonical representation, I believe that graphical notations that are more specific and directly represent biological processes are favored, owing to the familiarity of the visualization and capability to cope with diverse biological processes with compact representation. Nevertheless, it is certain that there are groups of researchers who favor canonical representation such as Petri net notation, rather than specialized notations. Which notation is to be accepted as the most commonly used notation depends on user choice, particularly when researchers have to deal with truly large networks with detailed biological features.

Flow chart and timing chart

Two additional diagrams - the flow chart and the timing chart - may enhance readability of the system. The flow chart represents a series of biological events that we use to intuitively describe the process under consideration. It is similar to the event model in Cook's notation. Each node represents an intuitively labeled biological state or process at arbitrary level of abstraction. Each node, however, needs to be mapped onto the node, or nodes, in the process diagram to ground it to the molecular level. Although nodes in flow charts only describe intuitive landmarks, the mere breakdown of each node and arc means that you would not be able to reconstruct detailed molecular models. Instead they should be used to glue multiple processes, each of which are based on detailed molecular models, or to visually represent the state of the biological processes at the intuitive level. A timing chart is a graph that represents time-course changes of values involved in the model. Typically, this represents concentration levels and activation levels of protein involved.

Conclusions

In this paper, a series of improvements and additions to an existing set of diagrams for biochemical reactions and gene regulations were proposed. MIM was used as the starting point of discussions and several improvements have been made. In addition, the process diagram has been defined as a way to intuitively visualize temporal sequence of specific reactions that cannot be easily captured by MIM. Proposals made in this article are by no means exhaustive and complete. There are several additional proposals and modifications that have to be made to further improve useful and solidly defined graphical notation of biological systems. It is expected that a standard for graphical representation of biological systems is formed soon, rather than later, and proposals in this article contribute to such efforts. The software tool 'CellDesigner' has been developed and released by the Systems Biology Institute to support notations described in this article. The current version supports part of the process diagram and is available from http://www.systems-biology.org/. Support for the improved version of MIM will be made in version 2.0, expected to be released in early 2004.

Defining graphical notation for biological networks is a formidable challenge. It is unfortunate that the value and difficulties of such efforts are generally under-estimated today. In the long run, however, establishment of standard notation will seriously impact biological research by fostering rigid and speedy description and dissemination of knowledge, as well as enabling high-level analysis of large-scale biological networks within the community. Recently, a forum has been created within the Systems Biology Markup Language (SBML [16]) community for creating standard graphical representation. With numbers of iterative release and community feedback, it is expected that such efforts will converge into widely acceptable graphical notation.

Acknowledgements

The author wishes to thank ERATO Project members for their valuable comments, especially Andy Finny, Mike Hucka, Hamid Bolouri, Harbart Sauro, Mineo Morohashi, Akira Funahashi, Noriko Hiroi and Torbjorn Nording. The author also wishes to thank anonymous reviewers for their constructive and informative comments.

This research is in part supported by the Rice Genome and Simulation Project (Ministry of Agriculture, Japanese government), the international standard formation program (NEDO, Ministry of Economy, Trade, and Industries), and a special coordinate fund for Keio University (Ministry of Education, Culture, Sports, Science, and Technology).

References

1 Kohn, K.W. (1999) Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Mol. Biol. Cell* 10, 2703–2734

2 Kohn, K. (2001) Molecular interaction maps as information organizers and simulation guides. *Chaos* 11, 84–97

3 Cook, D.L. *et al.* (2001) A basis for a visual language for describing, archiving and analyzing functional models of complex biological systems. *Genome Biol.* 2, RESEARCH0012

4 Pirson, I. *et al.* (2000) The visual display of regulatory information and networks. *Trends Cell Biol.* 10, 404–408

5 Maimon, R. and Browning, S. (2000) *Diagrammatic Notation and Computational Structure of Gene Networks*. In *Proceedings of the Second International Conference on Systems Biology,* Pasadena, CA, USA

6 Davidson, E.H. *et al.* (2002) A genomic regulatory network for development. *Science* 295, 1669–1678

7 Yuh, C.H. *et al.* (1998) Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science* 279, 1896–1902

8 Alberts, B. *et al.* (2002) *Molecular Biology of The Cell* (4th edn), Garland Science

9 Berg, J. *et al.* (2002) *Biochemistry*, W.H.Freeman and Company

10 Price, N. and Stevens, L. (1999) *Fundamentals of Enzymology,* Oxford: Oxford University Press

11 Voit, E. (2000) *Computational Analysis of Biochemical Systems,* Cambridge University Press

12 Petri, C.A. (1962) *Kommunikation mit Automaten*. In *Institut für Instrumentelle Mathematik,* Bonn, Germany

13 Matsuno, H. *et al.* (2000) Hybrid Petri net representation of gene regulatory network. *Pac. Symp. Biocomput.* 1, 341–352

14 Peleg, M. *et al.* (2002) Modelling biological processes using workflow and Petri Net models. *Bioinformatics* 18, 825–837

15 Matsuno, H. *et al.* (2003) Biopathways representation and simulation on hybrid functional Petri net. *In Silico Biol.* 3, 0032

16 Hucka, M. *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 19, 524–531