Using process diagrams for the graphical representation of biological networks

Hiroaki Kitano[1–4], Akira Funahashi[1,3,4], Yukiko Matsuoka[1,3] & Kanae Oda[1,4]

With the increased interest in understanding biological networks, such as protein-protein interaction networks and gene regulatory networks, methods for representing and communicating such networks in both human- and machine-readable form have become increasingly important. Although there has been significant progress in machine-readable representation of networks, as exemplified by the Systems Biology Mark-up Language (SBML) (http://www.sbml.org) issues in human-readable representation have been largely ignored. This article discusses human-readable diagrammatic representations and proposes a set of notations that enhances the formality and richness of the information represented. The process diagram is a fully state transition–based diagram that can be translated into machine-readable forms such as SBML in a straightforward way. It is supported by CellDesigner, a diagrammatic network editing software (http://www.celldesigner.org/), and has been used to represent a variety of networks of various sizes (from only a few components to several hundred components).

Drawing diagrams with nodes and arrows is the common approach for representing how proteins and genes interact, and papers frequently include such informal node-and-arrow diagrams. Although such diagrams are useful in providing an intuitive idea of how proteins and genes interact, the information contained in such diagrams is not precise because the syntax and symantics of the symbols used tend to be ambiguously defined. Often, arrows adopt multiple different meanings, so that correct interpretation of the diagram depends upon the knowledge of the reader. For example, **Figure 1a** shows a typical diagram often found in signal transduction papers. In this example, an arrow symbol could be interpreted four different ways: activation, translocation, dissociation of protein complex and residue modification. Correct interpretation of which biological process the arrow refers to depends entirely on the reader's knowledge. In general, such ambiguities and lack of information are not a major problem as long as the diagrams are small and represent genes, proteins and their local interactions. However, problems emerge when representing interactions within larger networks. Therefore, there is a need for diagrams that contain unambiguous process information in the symbols used and that can be transferred to standard machine-readable codes such as SBML for computational analysis[1].

Circuit schematic diagrams used in electronics are ideal examples of a graphical diagram. Engineers can reproduce the circuits drawn in the schematic diagrams without substantial additional information, because the diagrams are unambiguously defined, contain sufficient information and are based on well-accepted standards.

Kurt Kohn was the first to produce canonical representations for molecular interactions[2,3]; and other researchers have been working on alternative representations[4–8]. Unfortunately, none of the proposed schemes has been widely used for a variety reasons. For example, there is no software tool to create a Kohn Map efficiently, and this type of representation does not explicitly display temporal processes, which makes it difficult for readers to understand the sequence of events. Diagrammatic Cell Language (DCL) modifies Kohn's notation[9], but suffers from similar problems in that it does not explicitly display a temporal sequence of events and lacks publicly accessible documents and supporting software. Other notations have different shortcomings.

A successful diagram scheme must: (i) allow representation of diverse biological objects and interactions, (ii) be semantically and visually unambiguous, (iii) be able to incorporate notations, (iv) allow software tools to convert a graphically represented model into mathematical formulas for analysis and simulation, (v) have software support to draw the diagrams, and (vi) ensure that the community can freely use the notation scheme.

We have accumulated substantial experience in creating molecular interaction diagrams of various sizes, ranging from several components and interactions to several hundred components and interactions[10,11]. Whereas associations and combinatorial bindings of molecular species can be compactly described by an entity-relationship diagram (as exemplified by Kohn's diagram), temporal orders of reactions are made implicit so that intuitive understanding of the process of reactions is difficult. The process diagram explicitly represents the temporal order of reactions and states of molecules and complexes at the cost of an increased number of nodes and lines in the diagram. We have previously argued that either approach can be used, depending upon the purpose of the diagram, and both notations can maintain compatible information internally, but differ in visualization[7]. In our experience, however, a process diagram graphically representing state transitions of the molecules involved is more intuitively understandable than an entity-relationship diagram. This article describes in detail how process diagrams can be a vehicle for representing biological networks.

[1]The Systems Biology Institute, Suite 6A, M31 6-31-15 Jingumae, Shibuya, Tokyo, 150-0001 Japan. [2]Sony Computer Science Laboratories, Inc., 3-14-13 Higashi-gotanda, Shinagwa, Tokyo, 141-0022 Japan. [3]ERATO-SORST Kitano Symbiotic Systems Project, Japan Science and Technology Agency, Suite 6A, M31 6-31-15 Jingumae, Shibuya, Tokyo, 150-0001 Japan. [4]Department of Fundamental Science and Technology, Keio University, 3-14-1 Hiyoshi, Kohoku, Yokohama 223-8522 Japan. Correspondence should be addressed to H.K. (kitano@symbio.jst.go.jp).
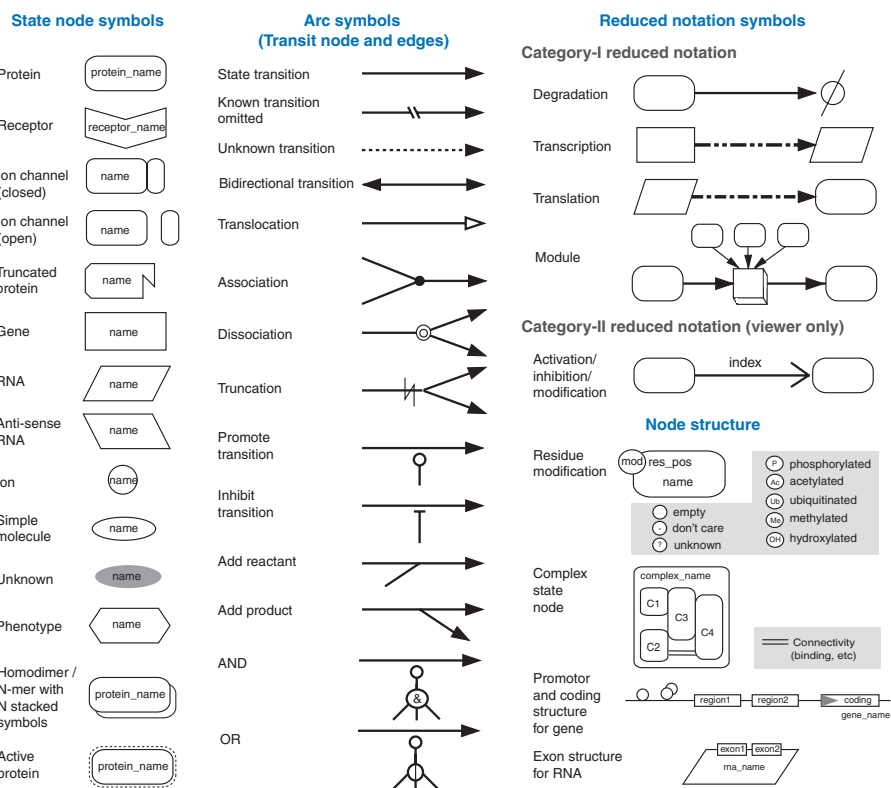
a

b

c
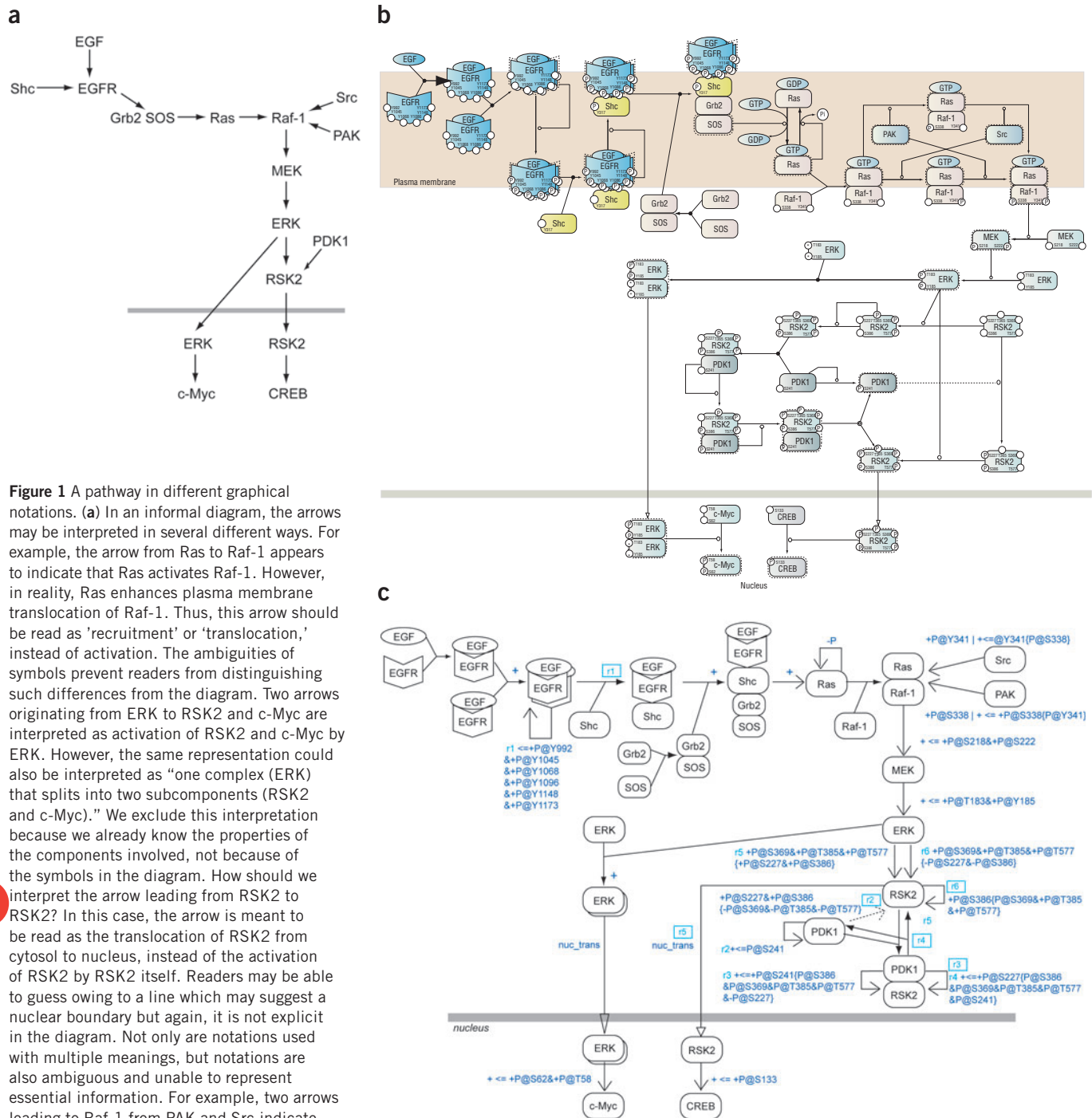
Figure 1 A pathway in different graphical notations. (**a**) In an informal diagram, the arrows may be interpreted in several different ways. For example, the arrow from Ras to Raf-1 appears to indicate that Ras activates Raf-1. However, in reality, Ras enhances plasma membrane translocation of Raf-1. Thus, this arrow should be read as 'recruitment' or 'translocation,' instead of activation. The ambiguities of symbols prevent readers from distinguishing such differences from the diagram. Two arrows originating from ERK to RSK2 and c-Myc are interpreted as activation of RSK2 and c-Myc by ERK. However, the same representation could also be interpreted as "one complex (ERK) that splits into two subcomponents (RSK2 and c-Myc)." We exclude this interpretation because we already know the properties of the components involved, not because of the symbols in the diagram. How should we interpret the arrow leading from RSK2 to RSK2? In this case, the arrow is meant to be read as the translocation of RSK2 from cytosol to nucleus, instead of the activation of RSK2 by RSK2 itself. Readers may be able to guess owing to a line which may suggest a nuclear boundary but again, it is not explicit in the diagram. Not only are notations used with multiple meanings, but notations are also ambiguous and unable to represent essential information. For example, two arrows leading to Raf-1 from PAK and Src indicate the activation of Raf-1 by these two kinases. However, it is unclear what the mechanisms are, which residues are phosphorylated, or which is the first modulator of Raf-1. Accompanying text can supplement missing information to explain such ambiguities, but in some cases the text might be more ambiguous than the diagrams. (**b**) In the process diagram, the meaning of symbols is defined more rigidly. An open arrow and a circle-headed line for Ras and Raf-1 indicates translocation of Raf-1 from the cytosol to plasma membrane (an open-arrow for translocation) promoted by Ras (circle-headed line for promoting the state transition). In addition, it indicates the specific activation mechanism of Raf-1 by Src and PAK. Raf-1 is fully activated via phosphorylation on both Tyr341 and Ser338 residues by Src and PAK, respectively. Each of the two arrows originating from ERK to RSK2 and c-Myc in **a** is represented in a very different way. The arrow heading to RSK2 is replaced by a circle-headed line, which indicates that RSK is phosphorylated by ERK, and subsequently stimulates its autophosphorylation. RSK2 is phosphorylated by two different processes with a specific sequence of events. The pathway from ERK to c-Myc is interpreted as a ERK homodimer formation and translocation to the nucleus, where homodimerized ERK activates c-Myc. When the reaction is described in this manner, an interpretation such as "one complex (ERK) split into two subcomponents (RSK2 and c-Myc)" is impossible. The translocation of RSK2 from cytosol to nucleus is shown by the open arrow and can be easily distinguished from state transition or catalysis. (**c**) An example of the process diagram with reduced notations. Each arrow for category-II reduced notation is associated with an index term that substitutes information that cannot be described graphically. This diagram lies between an informal diagram and a fully developed process diagram, but is much more informative and solidly defined than the informal diagram.

A process diagram is a state transition diagram with complex node structures. It consists of two classes of vertexes and edges. One class of vertex, called 'state node' (SN), represents the state of the entities involved in the biological process, such as proteins, small molecules, ions, genes and RNA. The other class, called 'transition node' (TN), represents modulations imposed on the reaction, such as catalysis, inhibition, association and dissociation. In a process diagram, different states of one molecular species are represented by different SNs. SNs that represent complexes are called complex SNs (CSNs), and there are two or more SNs as components of the node. There are two types of edges: edges from a state node to a transition node (ST-Edge) and edges from a transition node to a state node (TS-Edge). There are two types of TS-edges; one that represents state changes in the molecular species (represented by a closed arrow), and one that represents translocation of the molecule (represented by an open arrow). A reaction is represented as two or more state nodes connected by edges that are connected through a transition node. Each SN may have hierarchical internal structure defined as N-tree to represent members of a complex that are also SNs. Connectivity of internal nodes is defined by the connectivity matrix, which defines bindings among proteins that constitute a complex, as well as domains that constitute a protein. Each SN may have features that represent the modification state of residues as well as allosteric configurations. Mathematically, a network in the process diagram (PDN) is defined as PDN = (SN,TN,ST-Edge,TS-Edge)


Figure 2 Proposed set of symbols for representing biological networks with process diagrams. Symbols in the process diagrams consist of visual icons for state nodes and arcs. Each arc consists of a transit node and edges. Currently, there are four reduced notations that display simplified diagrammatic symbols. The category-I reduced notation can be used during editing of the network. The category-II reduced notation is limited to viewer software, and is not permitted during the editing process because of potential confusion that could arise from the implicit nature of state transition description.

where SN = $(sn_1, sn_2\ldots, sn_i)$, TN = $tn_1, tn_2,\ldots, tn_j)$, ST-Edge = SN × TN, TS-Edge = TN × SN, and $sn_i = (sn_j, sn_k, \ldots, sn_n : cm_i)$.

Each SN is assigned a graphical symbol that represents the type of entity the node represents, as well as graphical subscripts indicating features such as residue modification state (details are shown in **Supplementary Fig. 1** online). Each TN has a corresponding graphical symbol that represents the nature of the reaction. For example, promotion and inhibition of a state transition of a molecule are indicated by a circle-headed arrow and a bar-headed arrow, respectively. Although all state transitions are unidirectional, bidirectional reactions can be represented using two unidirectional state transitions with opposite directions. With this notation, the pathway shown in **Fig. 1a** is shown as **Fig. 1b** (full notation) or as **Fig. 1c** (with reduced notation). The symbols used to represent molecules and interactions are shown in **Figure 2**.

Using process diagram notations, the signal transduction pathway in **Figure 1a** would be written as shown in **Figure 1b**. There are several notable differences from conventional diagrams. First, unlike in conventional diagrams where an arrow generally means activation or inhibition, in the process diagram whether a molecule is active or not is represented by the state of the node (a simple example is shown in **Supplementary Fig. 2** online). Active nodes are visually distinguished with a dashed line around the node, but do not make a distinction between types of kinase activities. When a molecular species has different activation states, it should be represented by different nodes reflecting the different states of the molecule. It is primarily used as an optional visual aid for users, rather than to define the nature of activations. Promotion and inhibition of state transitions are represented as modifiers of state transition using a circle-headed arrow and a bar-headed arrow, respectively. An open arrow (arrow head not filled) indicates the translocation of a molecule that is a state transition in terms of the change in location of the molecule.

Second, the process diagram can visually represent the state of residues. The residue states are represented by circles on the rounded-corner-box associated with the type and location of the residue. It is important that residue modifications and other changes in the state of a protein are made visually explicit; modified states of the same protein will be treated as different entities in the simulation yet it must be clear that they are still the same protein.

A complex can be described as complex SNs that have an N-tree data structure with SNs as terminal nodes as well as connectivity matrices defining the connectivity among SNs. Graphically, this is represented as a nested rounded-corner box (**Fig. 2**). For example, NF-κB is a heterodimer of p65 and p50. The outer box can be named NF-κB referring to the complex, with two internal nodes representing p65 and p50, which are subunits of NF-κB. Contacting elements within the complex indicate that they are binding together, and a double solid line is used to represent the binding relationship among components within the complex when components cannot be directly aligned or where the use of double solid lines are more intelligible (**Supplementary Figs. 3** and **4** online). This formulation allows the representation of interactions
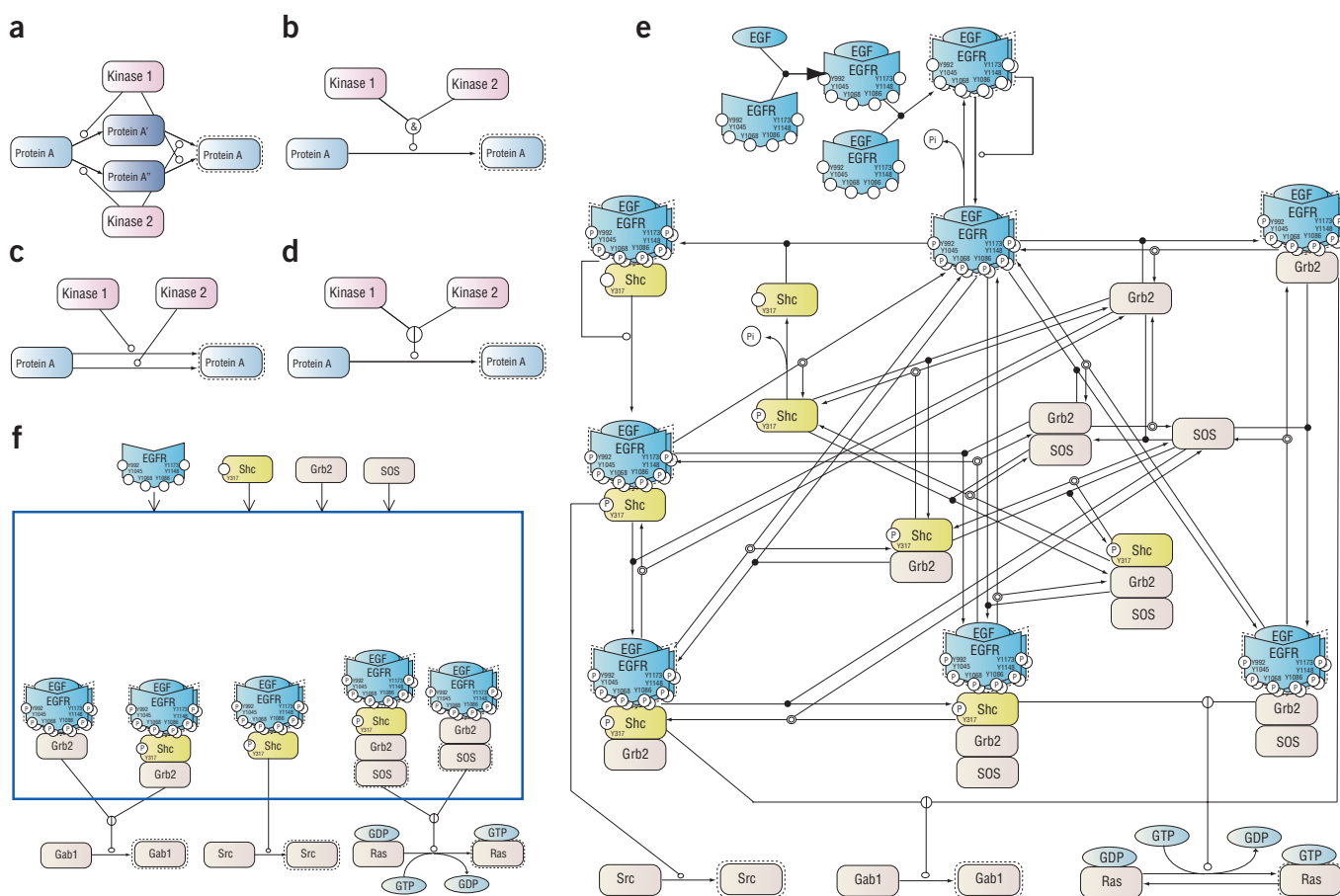
Figure 3 Representing combinatorial states. (**a**) Enumerating all state transition when two reactions takes place in random order. (**b,c**) A simplified view of the interactions shown in **a** (**b**), enumerating all transitions, any of which can transform the initial state into the final state (**c**). (**d**) A simplified form of the interaction shown in **c**. (**e**) A diagram that represents all combinatorial states of epidermal growth factor receptor (EGFR), Shc, Grb1 and SOS. Only three downstream interactions are shown here owing to space limitations. (**f**) A module representation of EGFR complex state transitions.

within a complex that has proteins in its components with kinase activity, or the representation of multiple binding and catalyzing domains of a protein (**Supplementary Fig. 4**).

The combinatorial explosion of states of molecules, binding combinations and multiplicities of interaction pathways makes the representation of biological processes very difficult. Consider the case of an unphosphorylated complex being transformed into a double phosphorylated complex mediated by two kinases, but the order of residue modifications is not a constraint. Such multiplicity of pathways can be handled by explicitly describing each set of interactions. The upper-right corner of **Figure 1b** shows an example of an unphosphorylated Ras-Raf1 complex being transformed into a doubled phosphorylated complex through interactions with Src and PAK via two routes. By the same token, RSK2 is phosphorylated via two distinct processes (**Fig. 1b**). Whereas the Ras-Raf1 subnetwork represents a case in which the order of interactions is not specified, the phosphorylation process of RSK2 has a specific order, such as phosphorylation of T385 and S389 by ERK (extracellular signal-regulated protein kinase), autophosphorylation of T577 and S386, and heterodimer formation with PDK1 and so forth.

When there are multiple intermediate states, whether residue modification states or binding states, the subnetwork representing this process could be extremely complex due to combinatorial explosion of possible states. There are four cases that must be considered.

First, in the case that every different state has significance in the given context of modeling or representation, every node has to be represented.

Second, in the case that only the initial and final state are important and a set of intermediate reactions can occur in random order but all intermediate reactions have to take place, then it can be represented by enumerating all intermediate states (**Fig. 3a**) or by using an AND logic symbol (**Fig. 3b**). When there must be a specific order in which reactions take place, then intermediate states after each reaction need to be represented.

Third, in the case that only the initial and final state are important and only one of the interactions is necessary for the transition, then it can be represented by drawing parallel state transition lines corresponding to each reaction (**Fig. 3c**), or by using an OR logic symbol (**Fig. 3d**).

Fourth, when the numbers of combinatorial states and associated state transitions are too large to be represented within the diagram, such a sub-network can be visually hidden as a 'module,' so that only a hexagonal box indicating a module is shown on the main diagram and detailed interactions can be described separately. This greatly reduces the visual complexity while details can be retained elsewhere. However, there are cases when a few states among a large number of combinatorial states have significance in reactions downstream or elsewhere, and only nodes representing these states can be visualized inside the module box, whereas all other nodes and state transition arrows can be visually

omitted. **Figure 3e** shows an example of a network for the epidermal growth factor receptor complex. **Figure 3f** illustrates how a module can be introduced to simplify the appearance when not all details are required. The module hides a number of intermediate states and complex interactions, and illustrates only specific intermediate states that affect other processes. The internal structure of the module is described separately. Although this combinatorial issue appears to be problematic, explicit enumeration of all possible states has to be done to ensure computational simulation. This is a fundamental trade-off between the process diagram and relationship diagram (such as Kohn's diagram). The relationship diagram can compactly represent combinatorial states, but temporal orders of reactions are implicit so that users have to follow lines and nodes to reconstruct the process. The process diagram represents a sequence of interactions as a state transition diagram, but all combinations may have to be enumerated where necessary. The use of modules significantly reduces the problem of readability that arises from the combinatorial explosion issue.

Transcription is described as the state transition of nucleotides into RNA, and translation is the transition of amino acids into protein. The process diagram enables detailed transcription and translation processes to be described (**Supplementary Fig. 5** online). However, in many cases, it is not necessary to describe detailed transcription and translation processes, so that a simplified notation that we call 'reduced notation' (**Fig. 2**) is often used.

The structure of a promoter region is represented as a rectangular box on the upper edge of the box that represents a sequence of DNA. By the same token, exons of RNA are represented as boxes on the upper edge of RNA symbols. For a simple description of transcription, genes are represented as simple rectangular boxes and transcription factors and other regulatory factors bind to the box.

Although interactions can be fully described using the state transition network, it is often too verbose and requires the description of excessive details that are not necessarily of interest. For example, ubiquitin-mediated degradation, transcription and translation may be simplified unless specific interactions of such processes are of central interest in the description. Thus, reduced notations can be used for these processes (**Fig. 2**).

There are two classes of reduced notations. The category-I reduced notation is a simplification of the visualization and model representation of intermediate processes, such as transcription, translation and degradation processes, when these processes are not of major concern. These notations can be used for efficiently editing biological network diagrams.

The category-II reduced notation represents molecular interactions that lead to activation, inhibition and state change of the protein. In some cases, simplified symbols are convenient for understanding the overall flow of interactions at the cost of precision and details. **Figure 4** gives a basic definition of the category-II reduced notation and some examples (details can be found in **Supplementary Fig. 6** online). The use of indexing to each arrow enables the effects and conditions for the interaction to be described. For example, an arrow from Raf1 to MEK with an index "+<= +P@S218&+P@S222", as seen in **Figure 1c**, means that "Raf1 phosphorylates S218 and S222 residues of MEK which causes MEK to be active." This index is important as it enables state transition to be represented while maintaining the style of a simple node and arrow diagram. Nevertheless, how activation and inhibition are triggered is not explicitly shown and the orders of state transition leading to change in activity state have to be carefully decoded by the reader. Unlike the category-I reduced notation, such as transcription, translation and degradation, which can be used within detailed process diagrams for the network model editing process, the category-II reduced notation for activation, inhibition and other modifications is used only for visualization of specific pathways where temporal

Syntax for index on category-II reduced notation
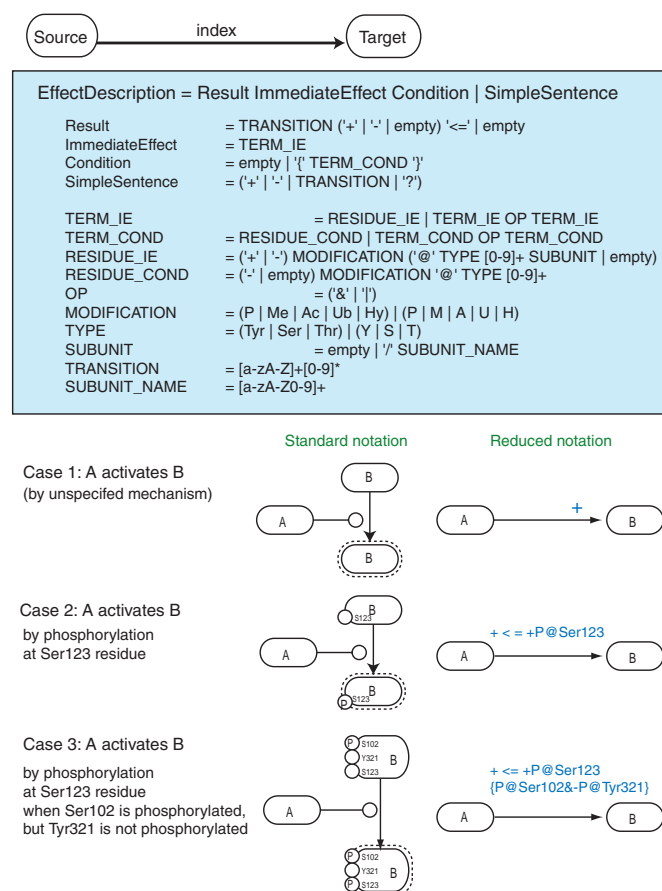
Figure 4 Syntax of index for category-II reduced notation and correspondence with the standard process diagram notation. The syntax is shown as a context-free grammar so that parser software can be easily built. The correspondence between regular process diagram notations and reduced notations illustrates how the index should be written.

orders of event are not critical. (Detail of limitations of reduced notation is described in **Supplementary Fig. 7** online.) On the contrary, the process diagram can be converted into a Kohn diagram and back (**Supplementary Fig. 8** online).

This article has discussed the benefits of the standardized canonical notation and described the process diagram as a basis of such notation. For such a diagram notation to be practical and to be accepted by the community, it is essential that software tools and data resources to be made available. CellDesigner (http://www.celldesigner.org/), a graphical editing software, has been developed to support visual editing of the network using the process diagram notation[10–12]. At this moment, CellDesigner supports most of the process diagram notation, and will fully implement the notation in the near future. Using the process diagram, a large-scale molecular interaction process map of ~600 components and interactions has been developed to demonstrate the scalability of the notation[11–13]. The pathway modules in PANTHER service by Applied Biosystems is an example of extensive use of the notation and CellDesigner software by a third party, and has been shown to be effective[14].

For the process diagram to be useful beyond graphical display it is essential that the diagram can be translated into machine-readable model representation language, such as SBML, in a straightforward

manner. This is trivial for the process diagram as each node and arc corresponds to species and reactions in SBML. Some of the reduced notation requires careful decoding of the diagram as it represents multiple steps in one reaction and different states of a molecule in one node. For computer simulation, each state of species and complex has to be distinguished and represented independently.

The graphical representation of biological networks is a topic that has been largely neglected, and its importance has only recently been recognized because of the growing need to understand large scale biological networks depicted by genome-wide analysis and other comprehensive measurements. A part of the notations proposed in this article has been used in the current version of CellDesigner, which is software for editing network models, to describe large scale models as well as small and medium scale models for numerical simulations by our group and others. A full set of notation shall be implemented in the future version of CellDesigner. The notation system is by no means complete and successive improvements need to be made based on feedback from various application cases. This article described version 1.0 of the process diagram notation, and a series of updates are anticipated to further enhance its capability and readability. However, the fact that a diagram containing several hundred nodes has been created with this method shows the potential of the proposed approach. The next possible step will be the formation of community to collectively define, improve and promote standardization of the graphical notation. Such an effort, named Systems Biology Graphical Notation (http://www.sbgn.org/), shall be synchronized with SBML and other standardization efforts to establish a widely acceptable and consistent framework for representation and communication of biological processes.

Note: Supplementary information is available on the Nature Biotechnology website.

1. Hucka, M. *et al.* The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**, 524–531 (2003).
2. Kohn, K.W. Molecular interaction maps as information organizers and simulation guides. *Chaos* **11**, 84–97 (2001).
3. Kohn, K.W. Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Mol. Biol. Cell* **10**, 2703–2734 (1999).
4. Maimon, R. & Browning, S. in *Proceedings of the Second International Conference on Systems Biology* (ed. Kitano, H.) 311–7 (Omnipress, Madison, WI, 2001).
5. Pirson, I. *et al.* The visual display of regulatory information and networks. *Trends Cell Biol.* **10**, 404–408 (2000).
6. Cook, D.L., Farley, J.F. & Tapscott, S.J. A basis for a visual language for describing, archiving and analyzing functional models of complex biological systems. *Genome Biol.* **2**, RESEARCH0012 (2001).
7. Kitano, H. A graphical notation for biochemical networks. *Biosilico* **1**, 169–176 (2003).
8. Aladjem, M.I. *et al.* Molecular interaction maps–a diagrammatic graphical language for bioregulatory networks. *Sci. STKE* **2004**, pe8 (2004).
9. Maimon, R. & Browning, S. in *Proceedings of the Second International Conference on Systems Biology* (Pasadena, California, November 5–7, 2001). (Eds. Yi, T.-M., Hucka, M,. Morohashi, M., & Kitano, H.) 311–317 (California Institute of Technology, Pasadena, 2001)
10. Funahashi, A., Tanimura, N., Morohashi, M. & Kitano, H. CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. *BIOSILICO*, **1**-159–162, (2003).
11. Oda, K. *et al.* Molecular interaction map of a macrophage. *AfCS Research Reports* **2**, 1–12 (2004).
12. Oda, K., Matsuoka, Y., Funahashi, A. & Kitano, H. A comprehensive pathway map of epidermal growth factor receptor signaling. *Molecular Systems Biology*, msb4100014–E1–E17 (2005).
13. Kitano, H. *et al.* Metabolic syndrome and robustness tradeoffs. *Diabetes* **53** (suppl. 3), S6–S15 (2004).
14. Mi, H. *et al.* The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res.* **33 (**Database Issue), D284–288 (2005).