SBGN Tutorial

An introduction to Systems Biology Graphical Notation

The goal of the Systems Biology Graphical Notation (SBGN) is to standardize the graphical/visual representation of essential biochemical and cellular processes studied in systems biology.

SBGN has excellent, detailed specification documents, which this tutorial will quote from extensively. However we will try to introduce the concepts in a slightly less dry fashion than the specification document.

SBGN defines a comprehensive set of symbols with precise semantics, together with detailed syntactic rules defining their use. It also describes the manner in which such graphical information should be interpreted. Standardizing graphical notations for describing biological interactions is an important step towards the efficient and accurate transmission of biological knowledge between different communities.

Traditionally, diagrams representing interactions among genes and molecules have been drawn in an informal manner, using simple unconstrained shapes and edges such as arrows. Until the development of SBGN, no standard agreed-upon convention existed defining exactly how to draw such diagrams in a way that helps readers interpret them consistently, correctly, and unambiguously. By standardizing the visual notation, SBGN can serve as a bridge between different communities such as computational and experimental biologists, and even more broadly in education and publishing.

For SBGN to be successful, it must satisfy a majority of technical and practical needs, and must be embraced by the community of researchers in biology. With regards to the technical and practical aspects, a successful visual language must meet at least the following goals:

1. Allow the representation of diverse biological objects and interactions;
2. Be semantically and visually unambiguous;

3. Allow implementation in software that can aid the drawing and verification of diagrams;
4. Have semantics that are sufficiently well defined that software tools can convert graphical models into formal models, suitable for analysis if not for simulation;
5. Be unrestricted in use and distribution, so that the entire community can freely use the notation without encumbrance or fear of intellectual property infractions.
6. Be comprehensible without software, for example in printed material.

Three types of SBGN diagrams

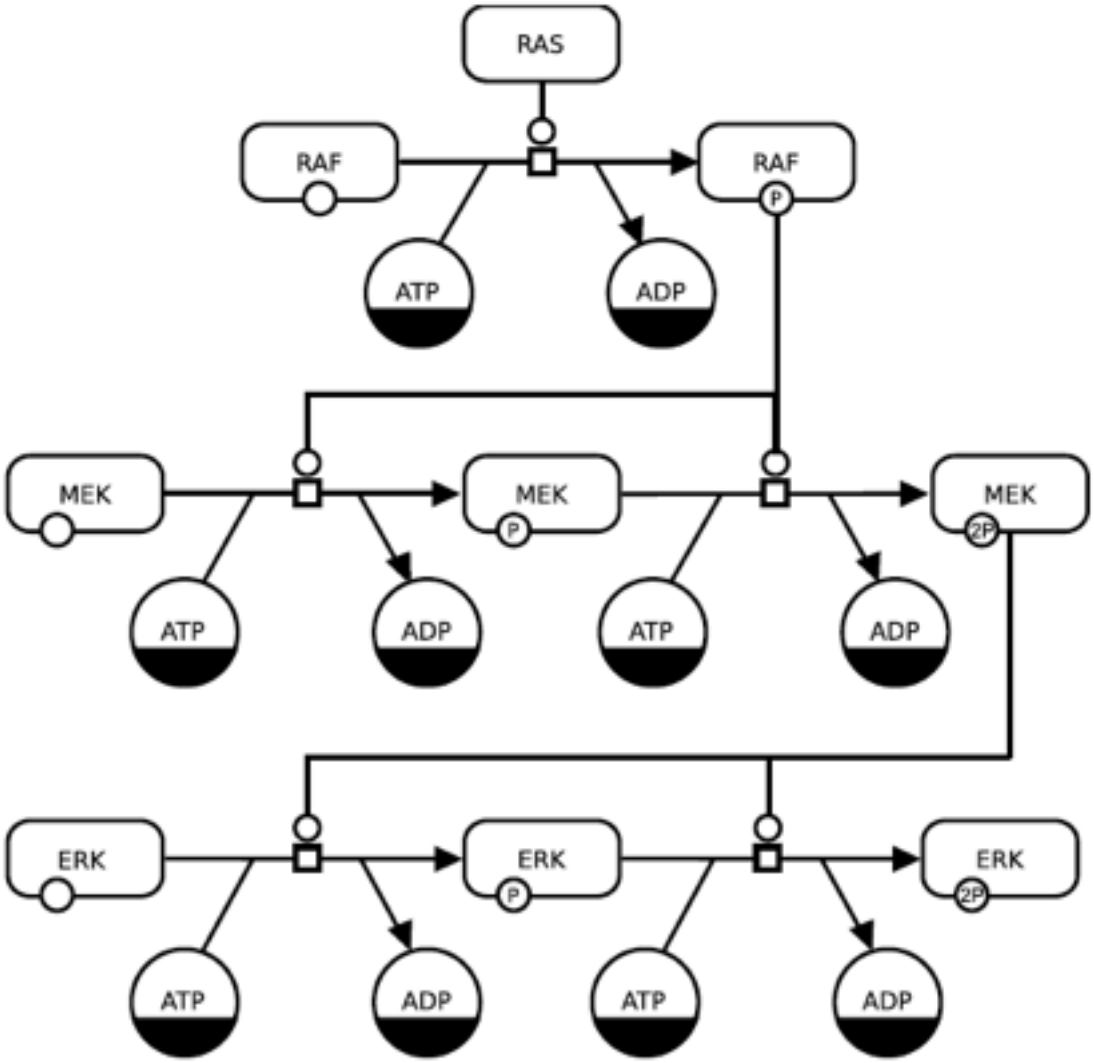
It is extremely difficult to represent all of the important parts of a biochemical network in a single diagram.
In SBGN, the term entity is mentioned many times. An entity refers to any biological thing # a compartment, compound, or macromolecule.
Three types of diagram notations are possible within the SBGN framework:

1. **Process Description diagrams (PD)** - causal sequences of processes and results. Although SBGN is not a modelling language, some of the concepts of this notation are derived from modelling. Of the three languages, this is most suited for conversion to a modelling description.
2. **Activity flow (AF) diagrams** # flux of information going from one entity to another.
3. **Entity-relationship (ER) diagrams** # the totality of interactions between species regardless of temporal sequence. This notation is designed for representing an independent set of known facts about a particular .

The specification for the Process Description was published first, and has the greatest software support so far.
This tutorial will focus mainly on Process Description notation.

The Process Description Notation

Below is an example diagram of the MAP kinase cascade, depicted in SBGN-PD:



There are symbols for macromolecules and simple compounds, and various types of connecting arcs, for **catalysis**, **consumption** and **production**. Macromolecules can contain *state labels* to convey notions of activity or post-translational modification. Entities that are repeated on the diagram are labeled with black interiors called *clone markers*. This signifies that the element appears more than once for the sake of visual clarity, rather than being a distinct component of the pathway.

The main purpose of the Process Description notation is to indicate change, showing how different entities are transformed by various processes.

The main types of visual features, or *glyphs* are:

1. **Entity Pool Nodes**, representing populations of things.
2. **Process nodes**, representing transformation.
3. **Connecting arcs**, linking transformations with things.
4. **Logical operators**, representing Boolean logic.
5. **Containment nodes**, representing physical restriction.

Restricted vocabularies

The Systems Biology Ontology (SBO) is used to provide a restricted vocabulary of terms for use in SBGN diagrams. For example, several common *material types* of entity are:

Description	Code	SBO Id
Non-macromolecular ion	mt:ion	SBO:0000327
Non-macromolecular radical	mt:rad	SBO:0000328
Ribonucleic acid	mt:rna	SBO:0000250
Protein	mt:prot	SBO:0000297
Polysaccharide	mt:psac	SBO:0000249

Within a diagram, an entity can be labeled with a *conceptual type* which indicates its role in the pathway.
For example:

Description	Code	SBO Id
Gene	ct:gene	SBO:0000243
Transcription start site	ct:tss	SBO:0000329
Gene coding region	ct:coding	SBO:0000335
Gene regulatory region	ct:grr	SBO:0000369
Messenger RNA	ct:mRNA	SBO:0000278

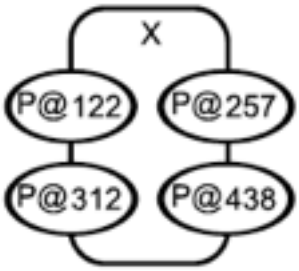
These can be used to annotate various entities, by inserting this information in a *Unit of Information*.

Entity nodes

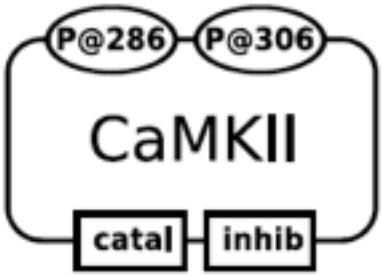
There are six types of entity pool node:

1. Unspecified entity
2. Macromolecule
3. Simple compound
4. Nucleic acid feature
5. Multimer
6. Complex

All of these entities can hold (or contain) clone markers, and units of information. In addition, macromolecules, nucleic acid features and multimers can hold *state glyphs*, which can indicate phenomena such as post-translational modifications.
For example:



indicates that protein X is phosphorylated at 4 sites.



shows that the protein is bi-phosphorylated, and has 2 functional domains (indicated in *Units of Information*) .

Container nodes

Container nodes are used to hold multiple entity nodes, to arbitrary levels of complexity.
The main container nodes are:

Complex - to represent macromolecular complexes
Compartment # to represent a physically restricted location in the cell.
Submap # indicating a complex process that should be mentioned, but whose details we are not concerned with in this particular diagram.

Other nodes

Source and *sink* nodes are used to denote an unspecified starting and ending points of a pathway, where we are not concerned with the details of the steps needed to create or destroy certain molecules in the pathway. For example, protein synthesis – we probably don't want to
include all the amino-acid incorporation steps.

Perturbations indicate some external stimulus on the system, e.g., light.

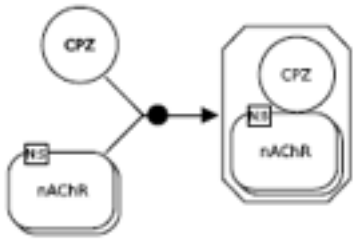
Phenotype is an externally visible alteration or biochemical process that is an outcome of the depicted pathway # for example, apoptosis or cell division.

Clone markers are used to label compounds that are repeated in several places in the diagram. These can include 'currency metabolites# as well as macromolecules. The idea is that by judiciously repeating nodes, the diagram is made clearer.

Process nodes

These nodes represent transformation. They are:

1. **Transition** : a generic, widely used glyph representing any transformation.
2. **Uncertain process** : can represent unknown or hypothetical processes.
3. **Omitted process** : a process whose details we are not concerned with in this pathway.
4. **Association** : the non-covalent binding of 2 or more entity nodes into a larger complex.
5. **Dissociation** : the breaking up of such as complex.



shows the association of a pentamer with a simple compound to form a complex.

Arcs

Arcs connect entities with processes. In SBGN, the following types of arc exist:

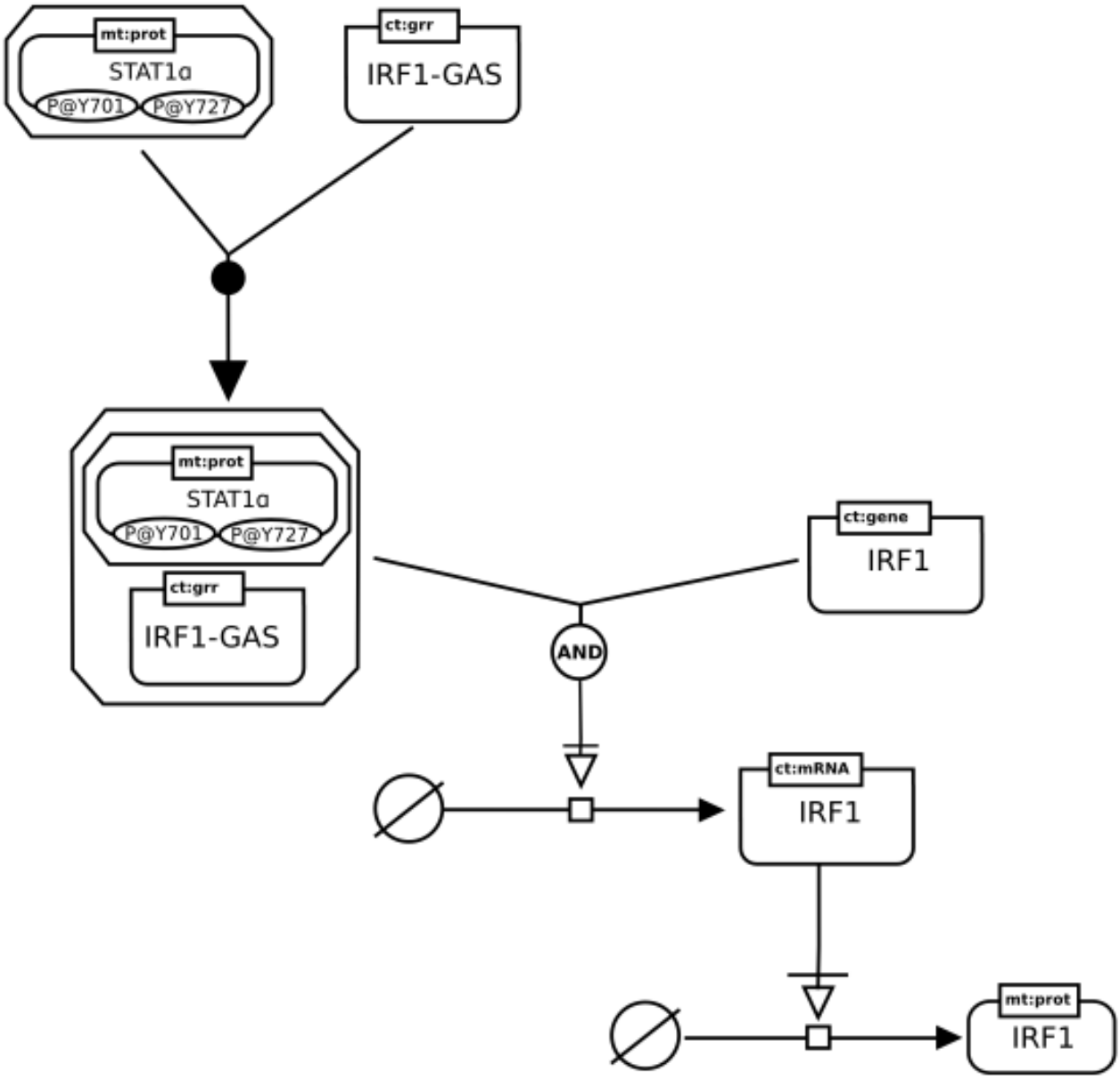
Consumption - links substrates to a reaction transition.
Production - -links a reaction transition to a product.
Modulation - indicates that an entity has either an unknown, or variable effect on a reaction process.
Stimulation - a positive influence on the basal rate of the process.
Catalysis # a special type of stimulation which acts by lowering the process#s activation energy.
Inhibition - a negative influence on the basal rate of the process.
Necessary stimulation - this is an #absolute# stimulation that is essential for the transition to occur # the basal rate is zero without such a stimulus.
Logic arc - connects logic gates to their input and output nodes.
Equivalence arc - indicates that an entity is equivalent to a tag.

Logic gates

And - all inputs are required for output
Or - any single input is sufficient for output
Not - a given input cannot produce the output

Exercises

This example diagram from the SBGN specification includes several concepts such as complexes, logic gates, units of information and association. Try to reproduce this diagram, using your software of choice.



Drawing hints # try to draw smaller components first, copying them and just giving them a different name. Create the entities first, then join them together, and align the shapes at the end.

If you have brought your own pathway diagram with you, you can try putting that into SBGN, using EPE or Cell Designer.

The entity relationship (ER) notation

The aim of this notation is to represent the totality of information about how entities (molecules) influence each other, and the relationships between them. In this notation, relationships are independent of each other. In this way diagrams can be drawn and edited based on specific conclusions from a scientific report.

In this notation, ‘entities’ are a more general concept than just chemical entities. Entities are considered as ‘sources of influence’, i..e, things that can modulate the activity of other components on the diagram. In SBGN-ER notation, then, there are the following entity types, which are explained more fully in the rest of the document:

1. **Entity interactor** - this can be any biomolecule,complex, heterodimer, or small molecule. No distinction is made between different molecule types, as is the case with the PD notation.
2. **Outcome interactor** - the outcome of an interaction or statement can be a source of influence. Outcomes are the consequences of an interaction or statement being true.
3. **Logical operators** combine statements, and have consequences depending on the evaluation of the logic.
4. **Perturbations** such as environmental changes, can also be a source of influence.

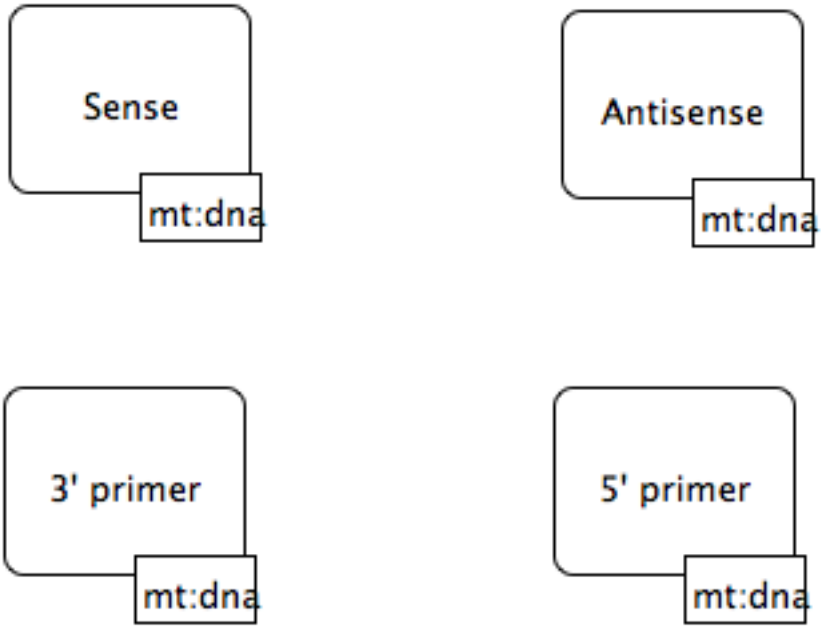
Relationships can be considered as rules which can be true or false. These are the relationships in SBGN-ER:

1. **Influences** - such as stimulation, modulation, inhibition etc
2. **Assignment statements** - depict setting the state of an entity to a particular value. For example, phosphorylation would be depicted as an assignment statement.
3. **Interaction statements** - depict any sort of binding, complex formation, dimerization etc.,
4. **Phenotypes** - represent the outcome of a statement that is outside the scope of the current pathway.

If a relationship is true, this can be the source of outcomes, which in turn can influence other entities.

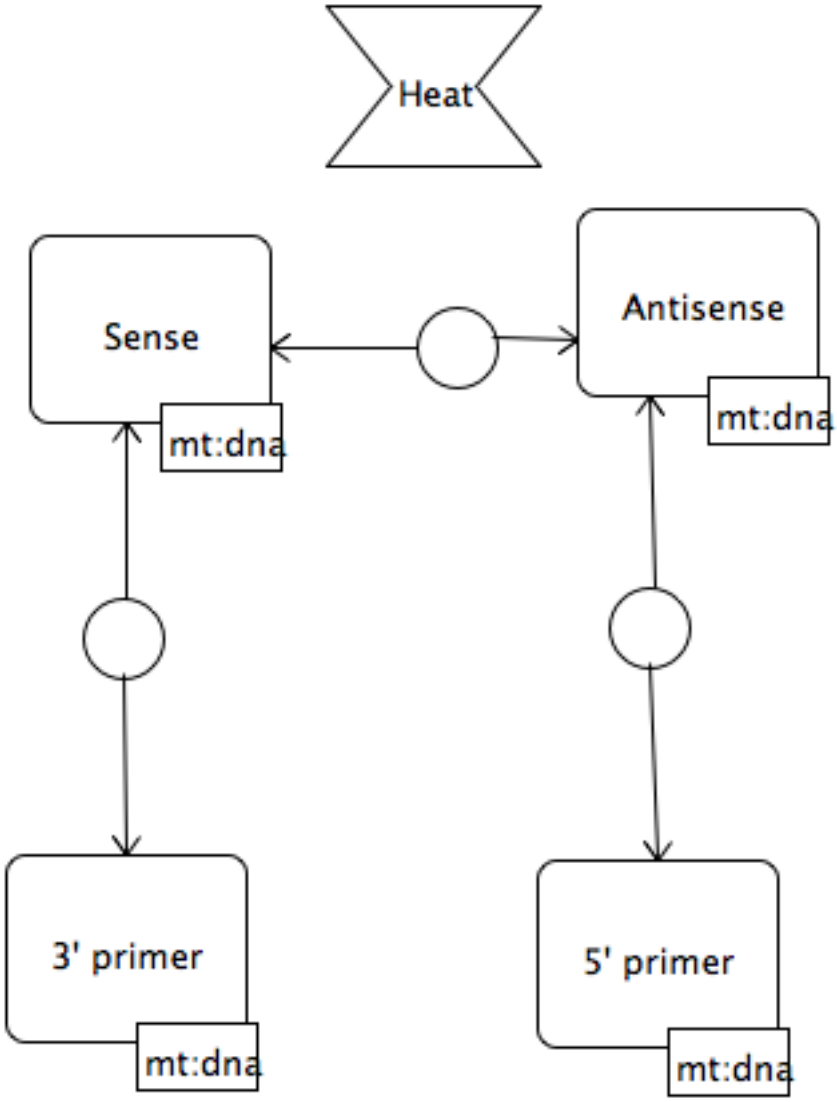
Let's illustrate some of the concepts involved by building a diagram describing the relationships between entities involved in the polymerase chain reaction (PCR).

Here are the *entity interactors* involved – the sense and anti-sense strands to be amplified, and the primers:

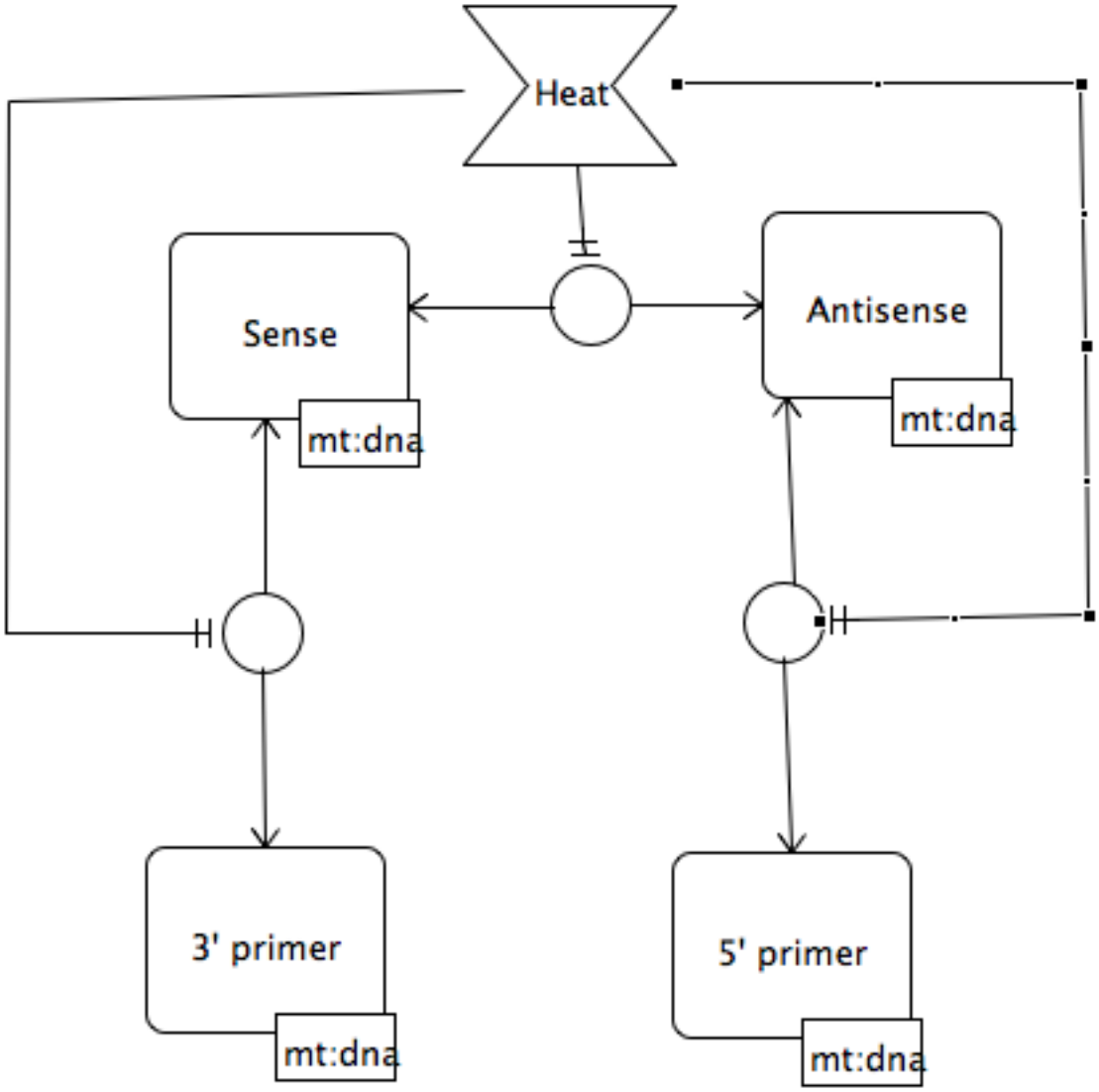


Unlike in the PD notation, all molecular entities have the same visual description – there is no distinction between macromolecules and metabolites, for example. Like PD notation, entities can have *Units of Information* associated with them, using the same vocabularies as well. In this example the entities are designated as DNA.

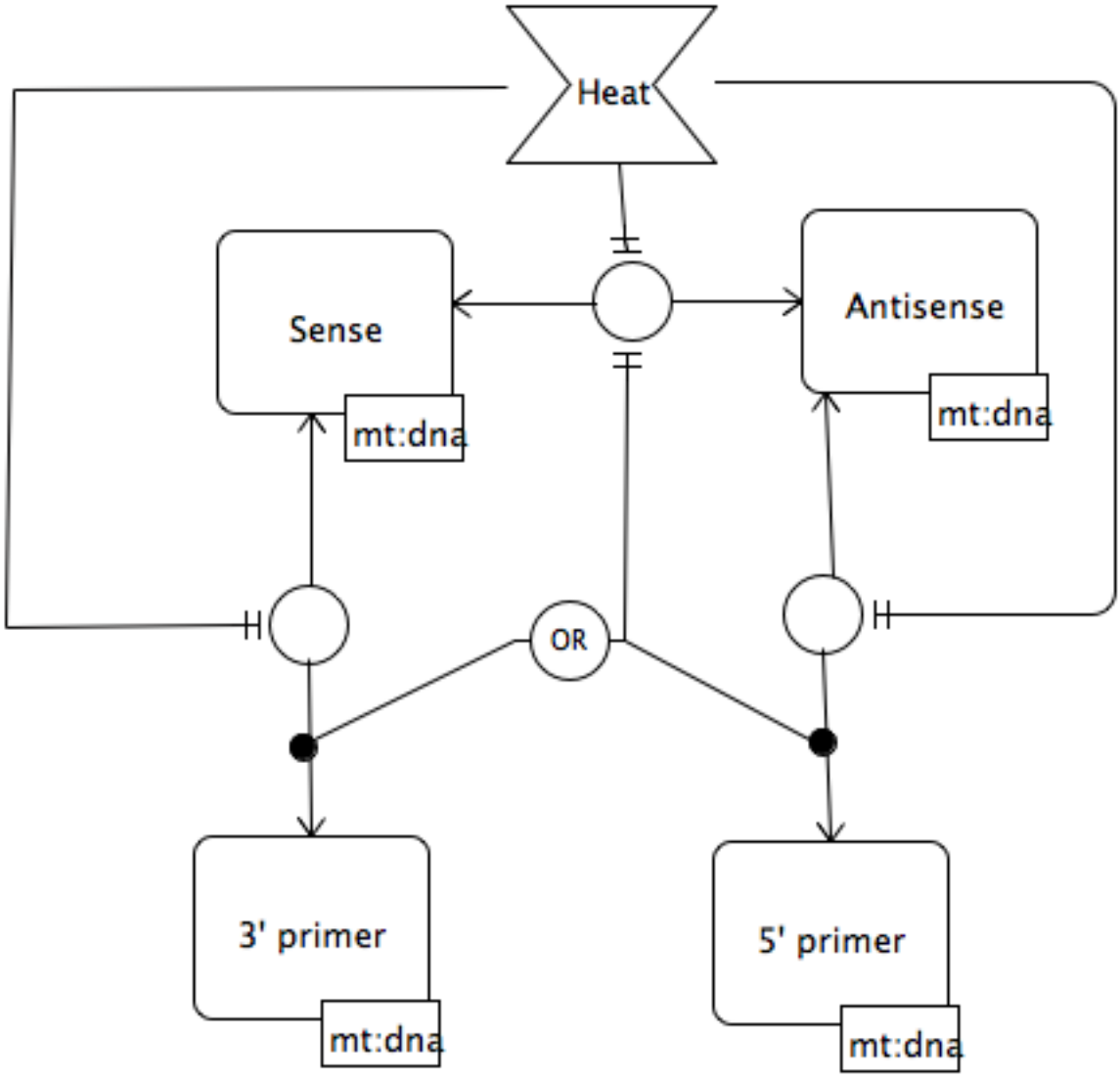
Now, we add some arcs to show some relationships between the entities – these arcs represent *interaction*. In this case the interaction represents a physical interaction (annealing). We've also added a *Perturbing agent* glyph, **Heat**, which is an entity, as it has an influence on other components
.



Heat obviously has a profound impact on a PCR reaction. ER notation can depict *influences* such as stimulation, inhibition, or modification. In the next picture we've added 3 arcs which indicate that heat totally inhibits the annealing of primers to the DNA strands.



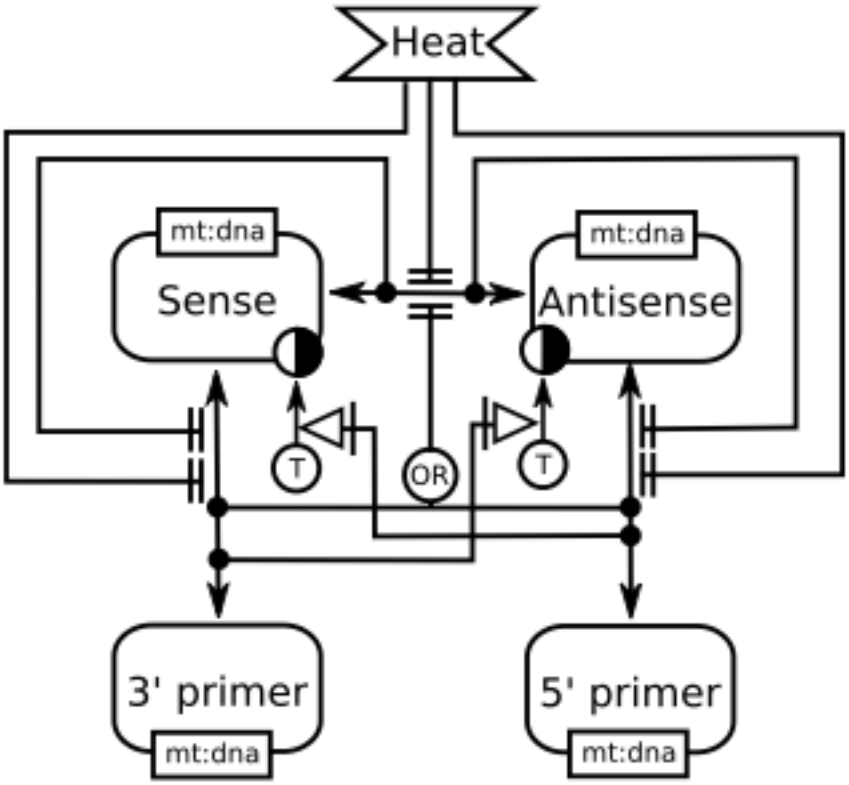
An important concept in ER diagrams is that of *outcomes*, or consequences. For example, in ER we can depict the outcome of an interaction or change of state of an entity. In the next diagram we depict a consequence of a primer binding its target – namely that the binding of the sense and anti-sense strands is inhibited. An *outcome* is depicted as a black dot. We also add an *or logic gate* to illustrate that either primer binding its target will inhibit the sense:anti-sense strand binding.



So, this new part of the diagram can be interpreted as “If either primer binds its target, then annealing of sense and anti-sense strands is inhibited”.

So far we’ve not addressed the notion of creation of elements. ER notation has a special symbol to represent existence:

Because ER attempts to present a logically consistent view, the concept of creation is thought of as “assigning truth to existence”. In this next diagram we’ve added these creation glyphs, and also two new arcs which originate from the outcome of the primers annealing to their targets. These can be read as: “If the primer binds its target, this has an outcome such that creation of a new strand is stimulated”. The state variables with the ‘T’ inside means ‘truth’.



New types of modulation.

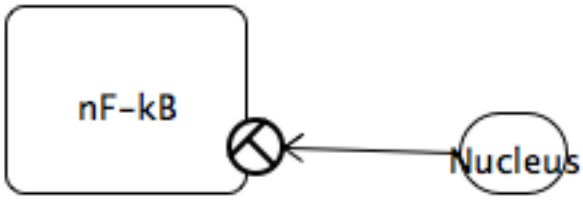
ER notation includes the following types of modulation as in PD notation – namely **Modulation**, **Stimulation**, **Inhibition**, **Necessary Stimulation**. However **Catalysis** is not included.

ER defines two new sorts of modulation arrow compared to PD: the **Absolute Inhibition** and the **Absolute Stimulation** influences.
Absolute Inhibition - this will inhibit a relationship regardless of other influences.
Absolute Stimulation - this will always trigger a relationship regardless of other influences.

Location

To describe location, the following glyph is used. It's a special sort of State Variable:

If we want to depict the fact that an entity has a particular location, in ER thinking this can be thought of as 'assigning location state to an entity'. E.g., the following means 'NF-kB is localized to the nucleus'.

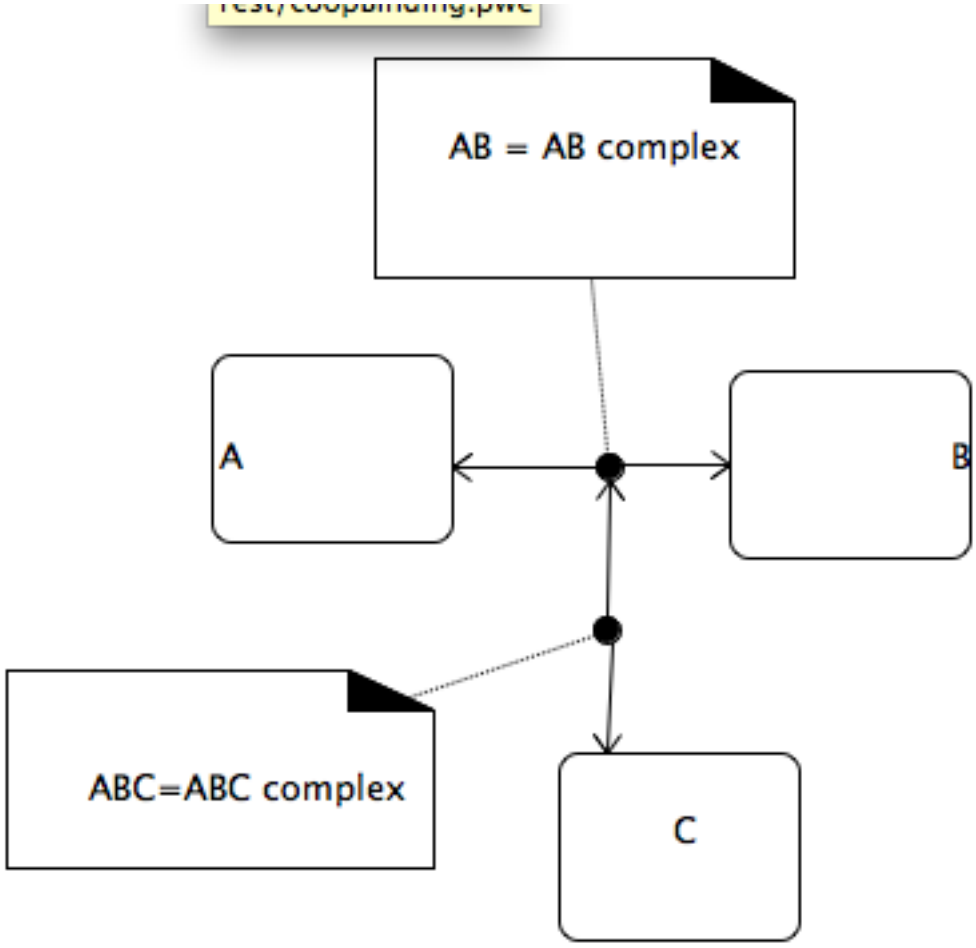


Examples of different sorts of binding in SBGN-ER notation.

ER diagrams are likely to be used extensively for depicting various sorts of interaction – below are some examples:

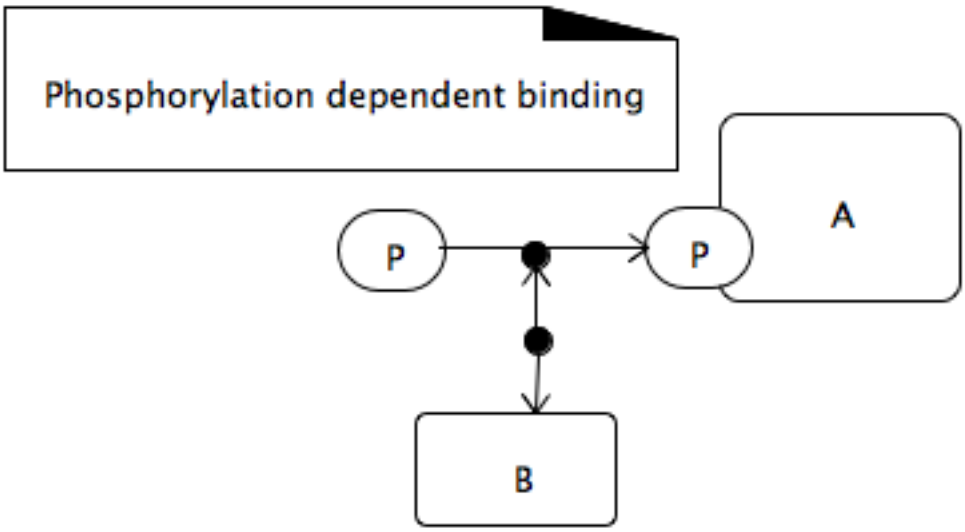
Multimer formation

In this diagram, the black dot between A and B represents the complex A:B, i.e., the outcome of A and B binding each other. A:B can subsequently interact with C to form complex A:B:C. The ability to connect outcomes in this way greatly reduces the combinatorial complexity of the diagrams.



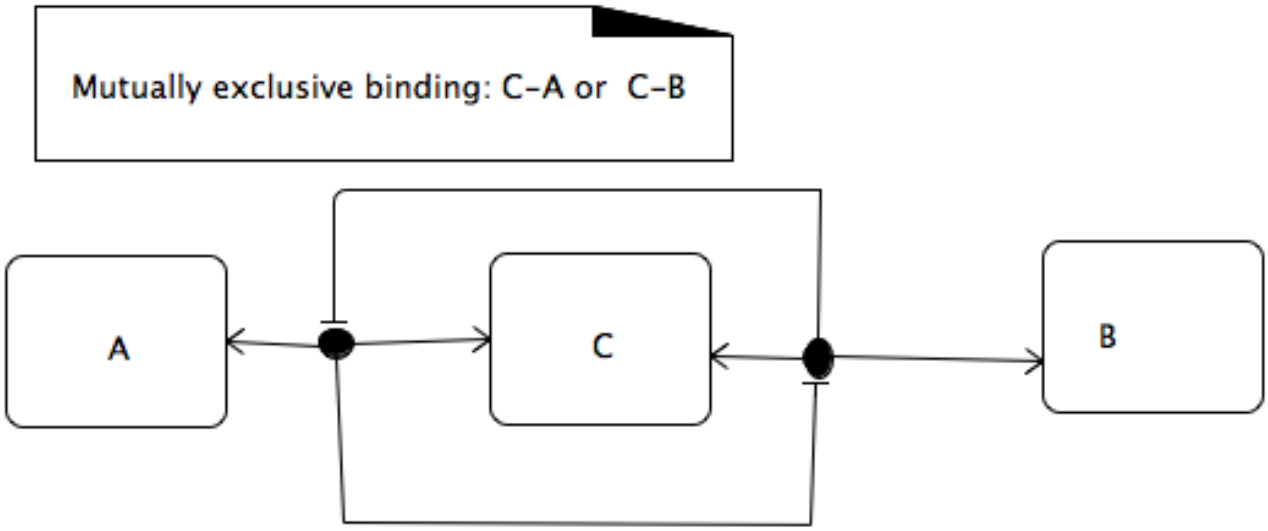
Phosphorylation dependent binding

In this diagram, B binds to the phosphorylated form of A. The lower black dot represents the complex $A_{phospho}$: B .



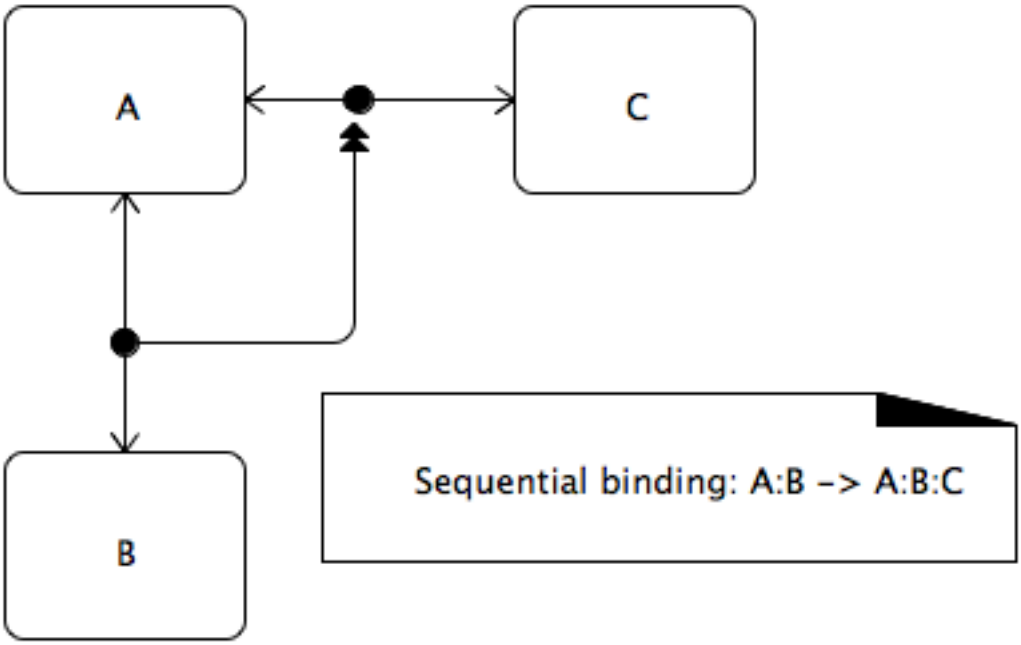
Mutually exclusive binding

In the diagram below, the formation of either C:A or C:B multimer inhibits the association of the third protein.



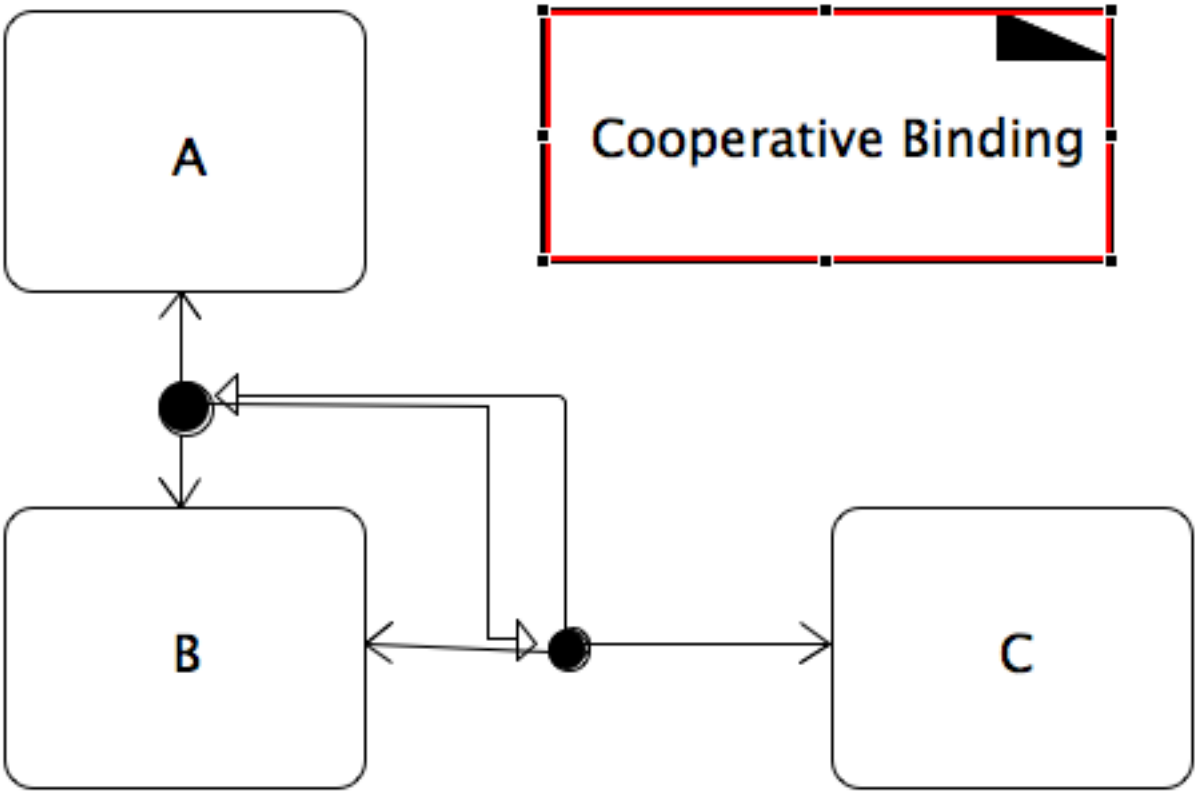
Sequential binding

In the diagram below, formation of A:B is necessary before C can join the complex.



Cooperative binding

In the diagram below, binding of either A:B or A:C stimulates the association of the 3rd protein.



The Activity Flow (AF) notation.

Introduction

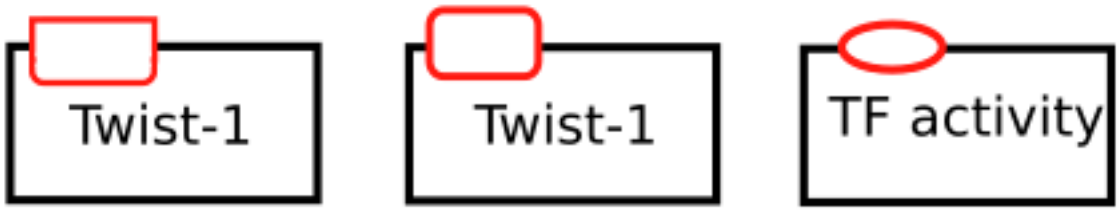
This notation is the least detailed of the three notations, in terms of the degree of information contained. Underlying detail of reaction mechanisms is not considered in this notation – it it were, then either PD or ER notations would be more appropriate.
This notation would be probably most familiar to bench biologists as a way to express functional behaviour without explicitly specifying the molecular details.

Glyphs in AF notation

Activity nodes

The core node is the 'Biological Activity'. This can represent a single biological entity, or a complex. Conversely, multiple activity nodes can be used to depict different activities of the same entity. A single glyph is used to represent an entity, with a *unit of Information* element used to give further detail about the node's type.

For example:



The shape of the *unit of information* glyphs is the same as that used in the Process Description notation - and thus can depict macromolecules, nucleic acid features, simple chemicals, unknown entities or complexes. In the illustration above we have 3 activities for Twist transcription factor – as a gene, macromolecule or unspecified.

In addition to the biological activity nodes, *Perturbation* and *phenotype* elements are also considered Activity Nodes as they can exert influence on the rest of the system.

Container nodes

As in PD notation, activity nodes can be contained inside *Compartments*. An activity in one compartment is considered distinct from that in another compartment, even if they have the same name.

Modulation arcs

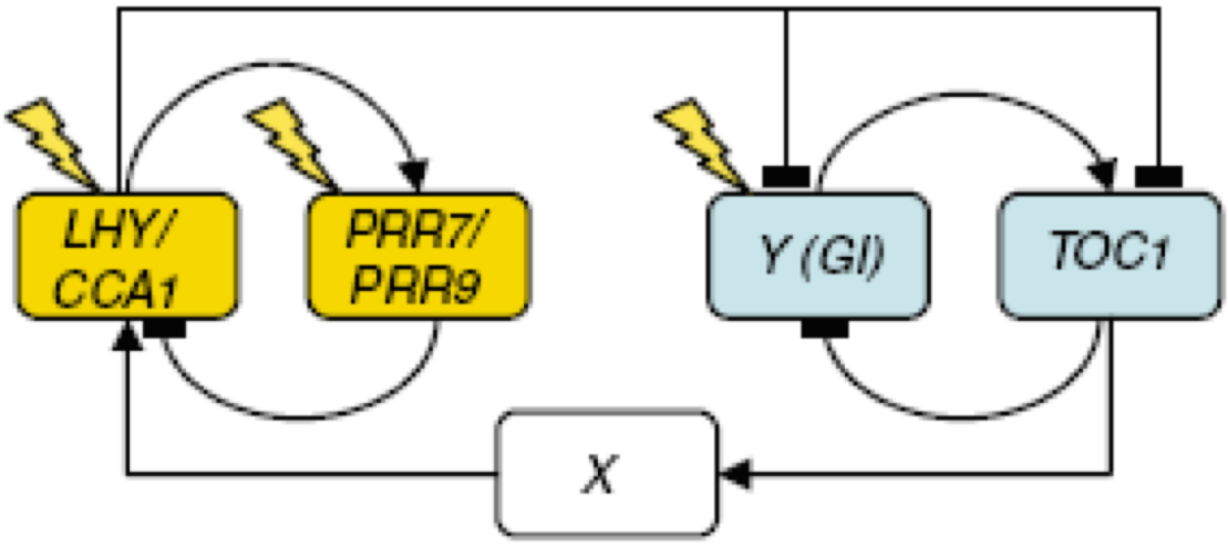
Activities influence each other by use of modulation arcs. These include *positive influence*, *negative influence*, *unknown influence* and *necessary influence*.

Logical operators

can be used to combine modulation arcs in logical combinations.

An example

Let's consider the three loop Circadian clock model:

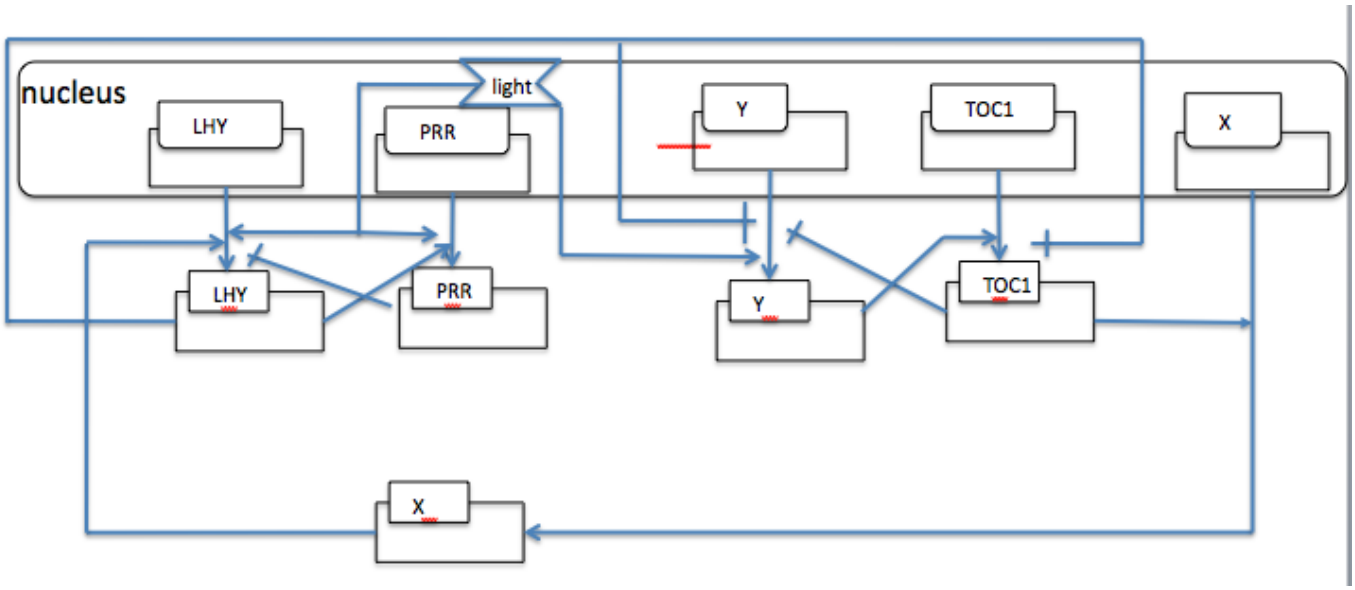


This diagram supposedly depicts 3 regulatory loops of the plant Circadian clock, which involves 4 transcription factors LHY, PRR and GI and X. .

1. First loop:
    1. Light stimulates LHY transcription. LHY protein returns to nucleus and stimulates PRR transcription.
    2. PRR #vely regulates LHY.
2. 2nd loop
    1. light stimulates GI production , which in turn activates TOC1.
    2. TOC1 #vely regulates GI and stimulates protein X
3. 3rd loop
    1. Protein X stimulates LHY production.
    2. LHY inhibits transcription of GI and TOC1

However, it's not very clear what is protein, what is gene, and where the reactions occur.

Below is this example redrawn in AF notation:



In this diagram, at the expense of slightly more complication, we have more definition of the processes involved, without introducing the concept of individual reactions and state transitions.

Comparison of the different notations

When would you want to use one notation over another?

The different notations present different views of biological knowledge.

PD notation would perhaps be most appropriate if some sort of modelling was intended for the pathway – there is more direct correspondence with say, SBML, than the other notations. More detailed knowledge is required of the set of reactions occurring, which is a pre-requisite for successful modelling.

ER notation is perhaps correct if one is trying to visualize conclusions drawn from the literature, or represent a set of non-kinetic knowledge.

AF notation could be used when trying to explain some biological functionality in a less detailed way, perhaps in an initial phase of capturing the essence of a biological pathway. It is simple enough so that it can be used manually, for example, drawing on a whiteboard, to convey ideas unambiguously.