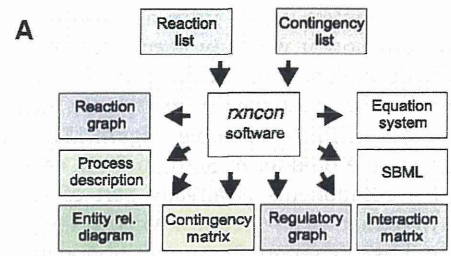
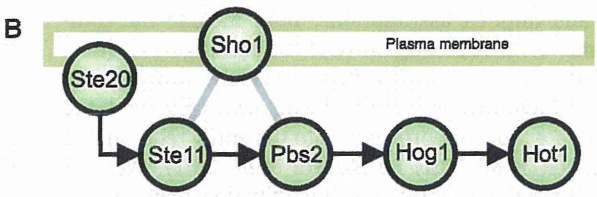


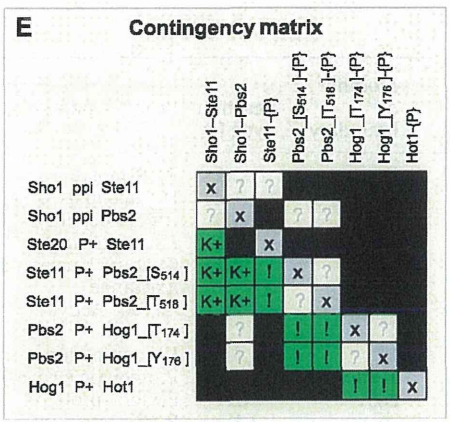


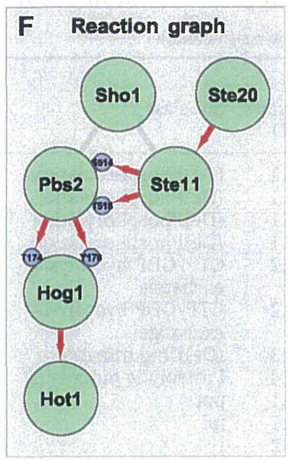
C **Reaction list**

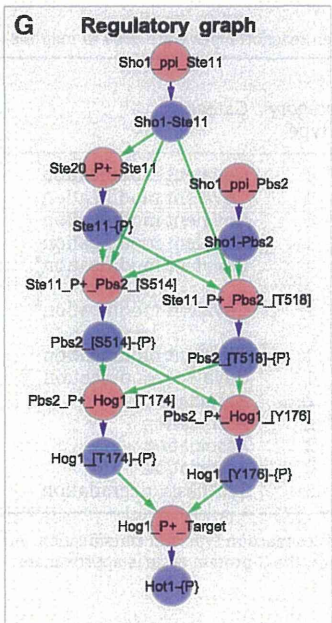
Comp. A	Re. type	Comp. B	Reaction	State
Sho1	ppi	Ste11	Sho1_ppi_Ste11	Sho1–Ste11
Sho1	ppi	Pbs2	Sho1_ppi_Pbs2	Sho1–Pbs2
Ste20	P+	Ste11	Ste20_P+_Ste11	Ste11-P
Ste11	P+	Pbs2_[S514]	Ste11_P+_Pbs2_[S514]	Pbs2_[S514]-P
Ste11	P+	Pbs2_[T518]	Ste11_P+_Pbs2_[T518]	Pbs2_[T518]-P
Pbs2	P+	Hog1_[T174]	Pbs2_P+_Hog1_[T174]	Hog1_[T174]-P
Pbs2	P+	Hog1_[Y176]	Pbs2_P+_Hog1_[Y176]	Hog1_[Y176]-P
Hog1	P+	Hot1	Hog1_P+_Hot1	Hot1-P

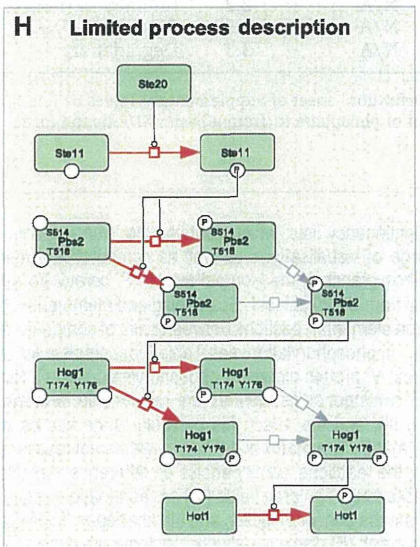
D **Contingency list**

Target	Cont.	Effector
Ste20_P+_Ste11	+	Sho1–Ste11
Ste11_P+_Pbs2_[S514]	+	Sho1–Pbs2
Ste11_P+_Pbs2_[S514]	+	Sho1–Ste11
Ste11_P+_Pbs2_[S514]	!	Ste11-P
Ste11_P+_Pbs2_[T518]	+	Sho1–Pbs2
Ste11_P+_Pbs2_[T518]	+	Sho1–Ste11
Ste11_P+_Pbs2_[T518]	!	Ste11-P
Pbs2_P+_Hog1_[T174]	!	Pbs2_[S514]-P
Pbs2_P+_Hog1_[T174]	!	Pbs2_[T518]-P
Pbs2_P+_Hog1_[Y176]	!	Pbs2_[S514]-P
Pbs2_P+_Hog1_[Y176]	!	Pbs2_[T518]-P
Hog1_P+_Hot1	!	Hog1_[T174]-P
Hog1_P+_Hot1	!	Hog1_[Y176]-P





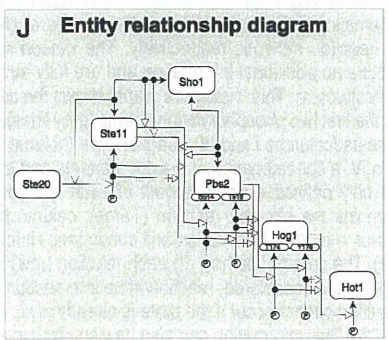




I **Interaction/distance matrix**

	Sho1	Ste20	Ste11	Pbs2	Hog1	Target
Sho1	-	-	ppi	ppi	-	-
Ste20	-	-	P+	-	-	-
Ste11	ppi	-	-	P+	-	-
Pbs2	ppi	-	-	-	P+	-
Hog1	-	-	-	-	-	P+
Hot1	-	-	-	-	-	-

	Sho1	Ste20	Ste11	Pbs2	Hog1	Target
Sho1	2	-	1	1	2	3
Ste20	2	-	1	2	3	4
Ste11	1	-	2	1	2	3
Pbs2	1	-	2	2	1	2
Hog1	-	-	-	-	-	1
Hot1	-	-	-	-	-	-



explosion and implicit assumptions. Both include the complete information about reactions (C1) and contingencies (C2). This data structure is also well suited for visualisation in entity relationship diagrams or extended contact maps, and the *rxncon* software tool supports export to the entity relationship format (Chylek *et al*, 2011; Le Novere *et al*, 2011). We also provide export to the reaction graph/activity flow diagram and the process description, though neither of these can fully and accurately represent the network as discussed below. Nevertheless, they all provide their unique advantages and can be automatically generated with the *rxncon* tool and the information in the reaction and contingency lists.

The contingency matrix integrates the information in the reaction and contingency lists (Figure 1E). The matrix is spanned by the reactions and their corresponding states (C1) and populated by the contingencies of reactions on states (C2). Each row corresponds to one elemental reaction and each column corresponds to one elemental state. The symbol in each reaction-state intersection specifies how that specific reaction depends on that specific state. Together, one row contains the complete set of rules a reaction follows, and hence describes how it works in every specific state. This is related to a dependency matrix (Yang *et al*, 2010), although the entries in the contingency matrix are more detailed and unambiguous. In the example (Figure 1E), the first row shows that (a) the binding of Sho1 to Ste11 cannot occur if either of the components is already part of such a dimer (column 1), (b) that we do not know whether the prior binding of Sho1 to Pbs2 (column 2) or phosphorylation of Ste11 (column 3) effects the Sho1–Ste11 binding and (c) that the other states appearing in the row are irrelevant for this specific binding reaction—as they do not describe properties of Sho1 or Ste11. The primary advantages of the contingency matrix are that it (1) allows a comprehensive documentation/visualisation of all reactions and dependencies within the network, (2) that it does so without requiring assumptions, (3) that it explicitly defines unknowns and hence gaps in our knowledge and (4) that the matrix constitutes a template from which mathematical models can be derived automatically (see below).

Table I Thirteen reaction types were used to map the MAP kinase network

Reaction	Category type	Category	Subclass ID	Subclass	Modifier or boundary	Reaction type ID	Reaction name
P+	1	Covalent modification	1.1	(De)Phosphorylation	P	1.1.1	Phosphorylation
P−	1	Covalent modification	1.1	(De)Phosphorylation	P	1.1.2	Dephosphorylation
AP	1	Covalent modification	1.1	(De)Phosphorylation	P	1.1.3	Autophosphorylation
PT	1	Covalent modification	1.1	(De)Phosphorylation	P	1.1.4	Phosphotransfer
GEF	1	Covalent modification[a]	1.2	GTP/GDP hydrolysis/ exchange	P	1.2.1	Guanine Nucleotide Exchange
GAP	1	Covalent modification[a]	1.2	GTP/GDP hydrolysis/ exchange	P	1.2.2	GTPase Activation
Ub+	1	Covalent modification	1.3	(De)Ubiquitination	Ub	1.3.1	Ubiquitination
CUT	1	Covalent modification	1.4	Proteolytic processing	Truncated	1.4	Proteolytic cleaveage
ppi	2	Association	2.1	ppi	N/A	2.1.1	Protein–protein interaction
ipi	2	Association	2.1	ipi		2.1.2	Intra-protein interaction
i	2	Association	2.2	i	N/A	2.2	Interaction (non-proteins)
BIND	2	Association	2.3	BIND	N/A	2.3	Binding to DNA
DEG	3	Synthesis/degradation	3.3	DEG	N/A	3.3	Degradation

The table indicates reaction type and classification. Additional details are provided in the 'Reaction Definition' sheet of Supplementary Tables S1 and S2.
[a]For convenience, the G-protein cycle is approximated as a covalent modification by addition/removal of phosphate to/from a basic, GDP-bound form.

Figure 1 Schematic representation of the data structure. (**A**) The input data are the reaction and contingency lists, which contains the 'what-aspects' and 'when-aspects' of the reaction network, respectively. The *rxncon* software uses these lists to create a range of visualisations as well as computational models. These conversions require no additional information and are fully automated. (**B**) A simplified version of the Sho1 branch of the Hog pathway in *S. cerevisiae* will be used to illustrate the data structure. This 'biologist's graph' shows the activating phosphorylation cascade (arrows) from Ste20 to Hot1. Scaffolding and membrane recruitment by Sho1 facilitates the first two phosphorylation events (grey lines). (**C**) The (simplified) reaction list defines the elemental reactions between pairs of components. It includes the two components (columns I and III), reaction type (column II; 'ppi' = protein–protein interaction, 'P+' = phosphorylation; see Table I for complete list of reactions), reaction (column IV, a concatenation of the components and the reaction type) and resultant state (column V; protein dimers or phosphorylated states). Note that each elemental state only defines a single aspect of each component's specific state. (**D**) The (simplified) contingency list defines the relationship between states and reactions. It contains the affected reaction (Target, column I), the influencing state (Effector, column III), and the effect this particular state has on that reaction (contingency, column II). (**E**) The reaction and contingency information is summarised in the contingency matrix. The matrix is defined by elemental reactions (rows) and states (columns). The cells define how (if) each reaction (row) is affected by each state (column); that is, the reactions' contingencies on different states. Note that only direct contingencies are considered; reaction/state intersections which do not share components are blacked out. The grey fields ('x') are automatic as states are binary and hence a reaction cannot occur if the state is already true. The green fields ('!'/'K+') are imported from the contingency list, and all other open fields are defined as unknown effect ('?'). This information can also be visualised in a number of graphical forms: The reaction graph (**F**) displays network topology with either components or their domains as functional units. The regulatory graph (**G**) combines the reaction and contingency information to display the causal relationship between the reactions in the network and provides a complete graphical representation of the knowledge compiled in the contingency matrix. The limited process description (**H**) displays the catalytic modifications in the signal-transduction network as state transitions with catalysts but without complex formation (compare Supplementary Figure S1). The interaction and distance matrices (**I**) provide a compact description of network topology and allow calculation of distances between nodes. Finally, the reaction and contingency data can be visualised as an entity relationship diagram (**J**). These visualisations and the equation system for this system, subsystem or your own favourite network defined in the same format can be automatically generated using the *rxncon* software.

The *reaction graph* displays a topological, directed reaction network (Figure 1F). It represents each entity as a single node and each relationship between a pair of entities as a single edge. Edges can be non-directional (e.g., protein – protein interaction), unidirectional (e.g., phosphorylation) or bidirectional (e.g., phosphotransfer). The full reaction graph displays the domains and residues involved in each reaction. The protein parts are independent nodes and defined as neighbours (proteins can have domains or residues, domains can have subdomains or residues, subdomains can have residues). The inclusion of domain information makes the reaction graph similar to the (extended) contact maps (Danos, 2007; Chylek *et al*, 2011). The reaction graph and contact maps are both purely topological and do not include any contextual information, in contrast to the extended contact map which, for example, may show that binding only occurs to phosphorylated residues. We also use a condensed variant that displays only the central node for each component and collapses multiple reactions of the same kind between a pair of components to a single edge, and hence corresponds closely to the activity flow diagram of SBGN (Supplementary Figure S1B; Le Novere *et al*, 2009). The advantages of the reaction graph are (1) the relative simplicity that makes it useful for visualisation of even large networks and (2) that it is suited for visualisation of large-scale data sets within the context of that network (see below).

The *regulatory graph* shows how information is conveyed through the network (Figure 1G). It improves on the reaction graph by including information on causality between the reactions in the network (C2 data). The regulatory graph shows the network's regulatory structure; that is, which reactions (via states) actually influence the rate of other reactions. It is a bipartite graph with the elemental reactions (red) and elemental states (blue) as nodes. Reaction-to-state edges simply show which reactions produce or consume which states. The state-to-reaction edges show which states (products of upstream reactions) affect the dynamics of which (downstream) reactions. These state-to-reaction edges correspond to the symbols in the contingency list, i.e., '!', 'K+', 'K−' or 'x'. The regulatory graph can easily be translated into an influence graph, which can be used for structural analysis of the network (Kaltenbach *et al*, 2011). In contrast to the influence graph or 'story' (Danos, 2007), the regulatory graph strictly separates the effects of reactions (production or destruction of states) and the modifiers (increase or decrease in reaction rates) via distinct edge types. Furthermore, only the (modified) elemental states are displayed and the (the unmodified) complementary source/target state is implicit. Hence, like in the 'stories', cyclic motifs only appear when there is a true feedback in the system. This visualises both the (possible) sequence of events and the feedbacks clearly. However, in contrast to the 'story', the regulatory graph is comprehensive and simultaneously visualises all possible paths or 'stories'. In this example (Figure 1G), the uppermost node pair corresponds to the reaction where Sho1 binds Ste11 (Sho_ppi_Ste11) and the resulting state Sho1-Ste11. The reaction-to-state edge linking these two nodes identifies Sho1-Ste11 as the product of this binding reaction. Note that the source states for this reaction are omitted (i.e., Sho1 not bound to Ste11 and Ste11 not bound to Sho1). The state-to-reaction edge from Sho1-Ste11 to Ste20_P+_Ste11 shows that the phosphorylation of Ste11 by Ste20 is enhanced in the Sho1-Ste11 complex. This reaction in turn produces the state Ste11-{P}, which is required for phosphorylation of Pbs2 on both Ser514 and Thr518. Hence, the information flow can be followed throughout the network as all edges are unidirectional. The main advantages of the regulatory graph are that it (1) allows a comprehensive documentation/visualisation of all reactions and contingencies within the network, (2) that it does so in a very compact format (3) without forcing non-supported assumptions, (4) that it can be used for structural analysis of the network and (5) that it clearly shows the information flow through the network.

Process descriptions are well established and allow visualisation of the information flow and mechanistic detail simultaneously (Kitano *et al*, 2005). They are excellent for representation of small networks which are completely known, but lack of data (of the right granularity) invariably lead to unsupported assumptions. In addition, these diagrams rapidly become very complex, generally forcing *ad hoc* reduction and additional implicit and unsupported assumptions. Therefore, process descriptions do not allow a complete description of the network with the stringency we require. However, the process description can be clear and easy to read, and we generate a limited version which excludes complex formation and hence avoids most of the combinatorial complexity. The difference is highlighted by the upper three nodes in the example (Figure 1H), where Ste20 phosphorylates Ste11. In contrast to full process description, the binding of Ste11 to Sho1, and how this binding would affect the phosphorylation, is not included (compare Supplementary Figure S1). The (limited) process description has several advantages: It (1) is intuitive to read and (2) defines in which internal state(s) an enzyme is active, its substrate and the exact target residue, which (3) conveys the information flow through the pathway, the enzyme–substrate relationships as well as the gaps in our understanding of these aspects.

The information can also be used to generate interaction matrices that specify which components react with which components. These can be rendered at several levels of detail ranging from a complete interaction matrix including protein domains and target residues that defines each interaction type, via condensed interaction matrices with only one row and column per protein that still contains reaction type information (Figure 1I, upper matrix), to numerical matrices that only include information on connection and directionality. We used the latter to calculate the distances within the network to generate a distance matrix (Figure 1I, lower matrix).

Finally, the *rxncon* tool provides export to entity relationship diagrams (Figure 1J). Like the regulatory graph, the entity relationship diagram displays reactions and contingencies separately and hence largely avoids the combinatorial complexity. The entity relationship diagram has the advantage of concentrating all information on a given protein around a central node, which works especially well for simple regulatory circuits. This emphasises the role of each component within the network, in contrast to the regulatory graph which emphasises the information flow through the network. The entity relationship diagram is generated automatically by the

rxncon software and visualised via Biographer (Biographer). In the same way, the *rxncon* software can be used to generate the contingency matrix, the reaction graphs, the regulatory graph, and, via BioNetGen (Blinov *et al*, 2004), the SBML file that constitute the basis for the process description. These generations are fully automated and hence the framework addresses the issue of (ii) automatic network visualisation without further assumptions and—in the case of the contingency matrix and regulatory graph—without any simplifications.

Generation of mathematical models

The contingency matrix is a template for automatic generation of mathematical models. Each elemental reaction corresponds to a basic (context-free) rule in a rule- or agent-based model (Table II), or, in other words, a set of rules that share a reaction centre (Chylek *et al*, 2011). All contextual constrains on an elemental reaction is defined in a single row in the contingency matrix, and this row defines the elemental reaction's implementation in the rule-based format. The basic rule suffices if there are no known modifiers of a particular elemental reaction (i.e., only '0' and '?' apart from the intersection with its own state(s) (which is always 'x' for a product state and '!' for a source state)). Every other contingency splits the expression in two rules; one when that elemental state is true and one when it is false. The number of rules needed only increases with the number of quantitative modifiers ('K+' and 'K−') as the qualitative modifiers sets the rate constant to zero in either the 'true' (for 'x') or false (for '!') case (see Supplementary information for details). The expansion to rules is fully defined in our data format and the *rxncon* software tool automatically generates the input file for the computational tool BioNetGen (Blinov *et al*, 2004). This file can be used for rule-based modelling, network-free simulation and creation of SBML files. The translation to and from the rule-based format is unambiguous in both directions, and we illustrate this with translation of a rule-based model of the pheromone response pathway (yeastpheromonemodel.org). This model contains lumped reactions which we translate to combinations of elemental reactions, resulting in a different equation structure but the same functionality given appropriate choice of rate constants (Supplementary Table S3). Furthermore, we cannot distinguish different identical proteins in, for example, homodimers, and can therefore not define strict *trans*-reactions within such dimers. Apart from these issues, we can reproduce the same model with only cosmetic/nomenclature differences (see Supplementary information for details). Hence, the framework addresses the issue of (iii) automatic model generation from the database of biological information.

Mapping the MAP kinase network

As a benchmark, we have used the presented framework and an extensive literature search to create a comprehensive map for the yeast MAP kinase network (Supplementary Table S1). Reactions have been defined with specific residues and domains whenever experimental support was sufficient. The degree of experimental evidence has been evaluated manually and individually for each entry, and references to primary research papers supporting each interaction have been included in the reaction and contingency lists (column 'PubMedIdentifier(s)'). We have used mechanistic data on reactions (C1) and a combination of mechanistic and genetic data on contingencies (C2) between reactions and reactants' states from primary research literature. The mapping is based solely on primary research papers and *de facto* shown data to ensure a high-quality network reconstruction. We chose to exclude almost all genetic data as indirect effects cannot be ruled out even in well-performed genetic screens. Finally, we decided not to include spatial data, as we found information especially on regulation of (re)localisation too sparse. To the best of our knowledge, we have eliminated all questionable information from the compiled data set, and convincing reactions lacking solid mechanistic evidence have been included but clearly and distinctly labelled.

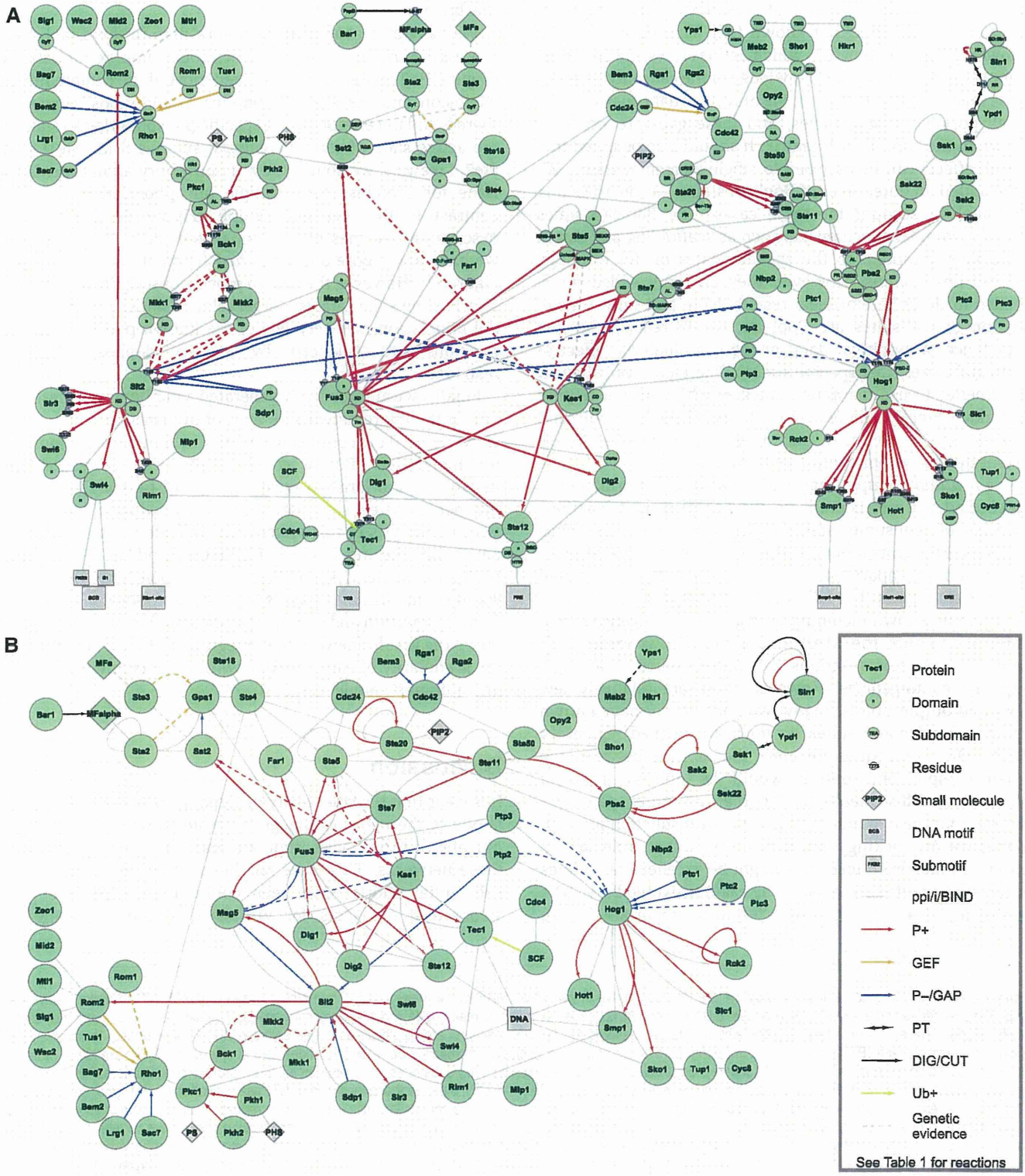
The MAP kinase network contains 84 components, 181 elementary states and 222 elementary reactions, corresponding to many hundreds of thousands of specific states. This network is large enough to be a severe challenge to the established visualisation and analysis methods. We did in fact fail to generate the complete state space and terminated the BioNetGen expansion after the first three iterations which generated 207, 1524 and 372 097 specific states, respectively. We use a range of graphical formats to visualise different aspects of this highly complex network. First, we display the network topology in the *reaction graphs* (Figure 2). These

Table II Implementation of elemental reactions in the rule-based format

Elemental reaction	BioNetGen rule implementation	
Interactions ("ppi", "i" or "bind")	A(B) + B(A) <-> A(B!1).B(A!1)	kf, kr
Intra-protein interactions (ipi)	A(A1,A2) <-> A(A1!1,A2!1)	kf
Phosphorylations (P+)	A + B(Psite~U) -> A + B(Psite~P)	kf
Autophosphorylations (AP)	A(Psite~U) -> A(Psite~P)	kf
Phosphotransfers (PT)	A(Psite~P) + B(Psite~U) <-> A(Psite~U) + B(Psite~P)	kf, kr
Dephosphorylations (P−)	A + B(Psite~P) -> A + B(Psite~U)	kf
Nucleotide exchanges (GEF)	A + B(GnP~U) -> A + B(GnP~P)	kf
Nuclease activations (GAP)	A + B(GnP~P) -> A + B(GnP~U)	kf
Ubiquitination (Ub+)	A + B(UBsite~U) -> A + B(UBsite~UB)	kf
Proteolytic cleavages (CUT)	A + B(Domain~U) -> A + B(Domain~truncated)	kf
Degradations (DEG)	A + B -> A	kf

The table displays how the different elemental reactions in Table I are translated to the rule-based format. See Supplementary information for additional details.

figures show that the number of characterised phosphorylation reactions vastly outnumbers that of characterised dephosphorylation reactions (68 to 16; Figure 2A), and that several well-established processes are only supported by genetic data (including the entire MAP kinase cascade below Pkc1; Figure 2B, dashed lines). The reaction graph also allows comparison between the established pathway architecture and the unbiased global protein – protein interaction studies and synthetic lethal networks (Figure 3A and B, respectively).



In the *contingency matrix* (Figure 4), we visualise the combined knowledge we have about the MAP kinase system (C1 and C2). The core matrix (red block of rows and blue block of columns) describe all the elemental reactions, elemental states and the (possible) contingencies of reaction on states. The black fields here show when there is no overlap between the components in the reactions and those defined in the states. Therefore, the matrix will always be sparsely populated. However, we also see that most of the remaining fields are grey; that is, effect not known ('?'). This means that our knowledge of reactions (C1; which defines rows and columns) is much stronger than our knowledge of the causality between these reactions (C2; the cells). We only have data on a minority of all possible contingencies, and these gaps are explicitly shown in the contingency matrix. It should also be noted that not all effects can be ascribed to single elemental states. We have added an outer layer of Boolean states (purple rows and columns) to account for these cases. The Boolean states describe complex mechanisms such as scaffolding and can in principle correspond to the specific states of, for example, process descriptions. However, they are only added when needed to describe empirical results. Note that only a small fraction of the states are Boolean, which reflects the low abundance of empirical data on the combinatorial effect of elemental states (i.e., specific states). Therefore, we believe it to be better to use mapping strategies which do not require such data. Finally, the matrix contains a layer of inputs and outputs (grey; columns and rows, respectively). These constitute the system's interface with the outside.

The *regulatory graph* (Figure 5) displays the information in the contingency matrix graphically, by showing how reactions produce or consume states, and how states influence reactions. This graph contains the full C1 and C2 information, and would fall apart without either. In fact, the isolated reaction–state pairs that fall outside the graph do so because they have no known incoming or outgoing contingencies. The graph shows that the MAP kinase network is rather well connected, as most reactions are indeed linked in a single graph by contingencies. However, there are relatively few input and output points; many reactions do not have known regulators and many states do not have defined regulatory effects. Only reaction–state pairs that appear between the system's input and output would be able to transmit information. This means either that all other pairs are irrelevant for the dynamics of the signal-transduction process, or that we are lacking information about their role in this process. In fact, such lose ends might be excellent candidates for targeted empirical analysis. One example would be Msb2's binding to Cdc42, which is reported to be important for the pseudohyphal differentiation pathway; raising the question of whether this binding is regulated in response to the stimuli that activate this part of the MAP kinase network. Another point that stands out is the almost complete lack of (documented) information exchange between pathways. The exception is the Sho branch of the Hog pathway, which is closely intertwined with the mating pathway, as both are activated by the shared MAP kinase kinase kinase Ste11 and parts of the cell polarity machinery.

We have also generated a network map in the established *process description* format, but without complex formations (Figure 6). This decision eliminated most of the combinatorial explosion and the need for implicit assumptions. However, there is still uncertainty in the specific phosphorylation state of the active state of certain catalysts, such as Ssk2, Ste11 and Ste7. Likewise, we do not know if phosphorylation order is an issue for proteins with multiple phosphorylation sites. In contrast to the regulatory graph (Figure 5), the process description becomes more complicated the more unknowns we have and Figure 6 is simplified (compare Supplementary Figure S2). However, the limited process description in Figure 6 clearly shows the catalyst–target relationships, and reinforces the impression that very few of the known phosphorylation reactions are balanced by known dephosphorylation reactions.

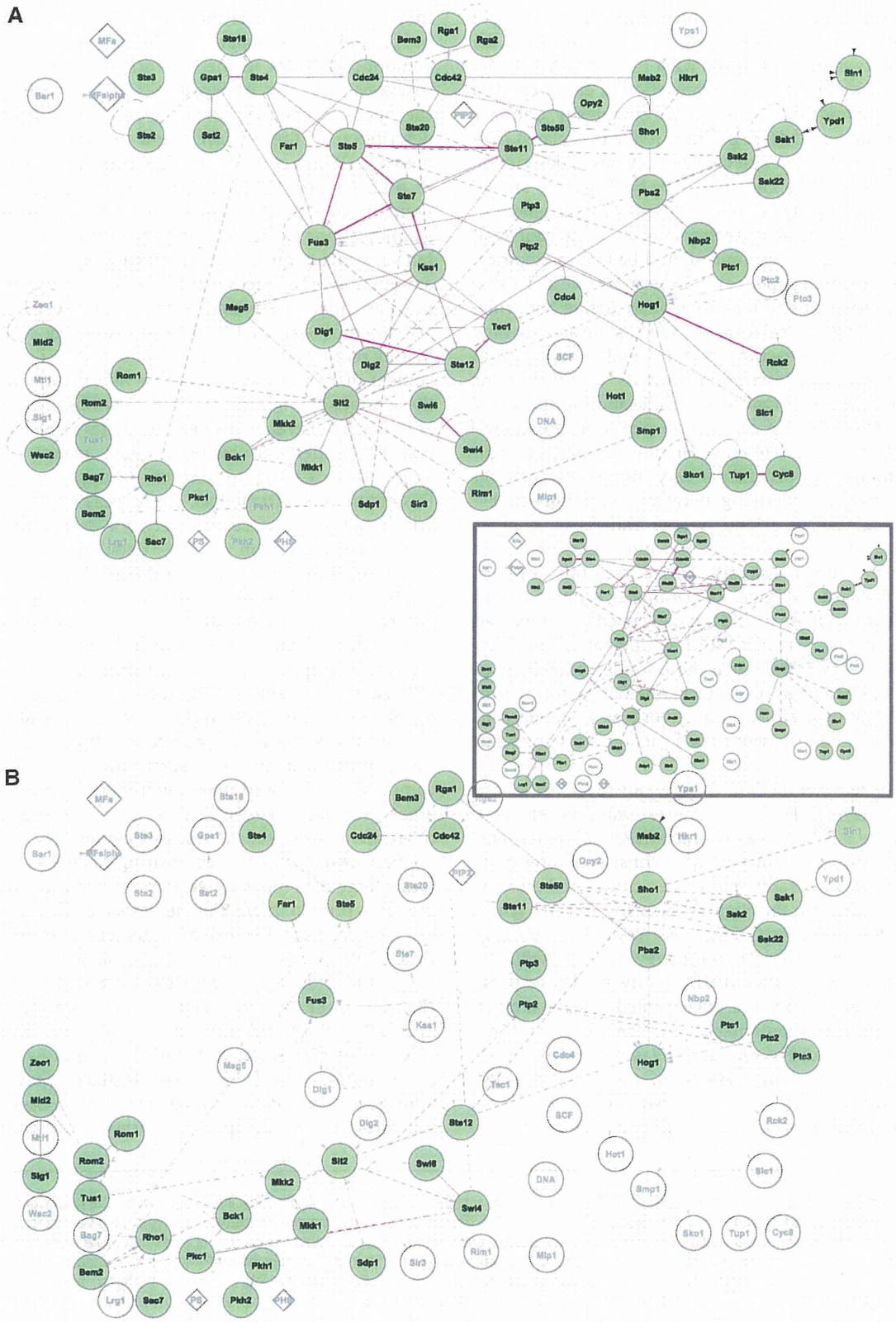
Finally, we automatically generated a mathematical description of the entire network as a proof of principle. The *rxncon* software used the contingency matrix to generate the input file for BioNetGen (Blinov *et al*, 2004). The corresponding network is too large to create but could be simulated with the network-free simulator NFSim (Sneddon *et al*, 2011). Further analysis of this system falls outside the scope of this paper, but the input file to BioNetGen and/or NFSim with trivial parameters is included as a supplement. Hence, a complete mathematical model can be automatically generated from the reaction and contingency data, and to our knowledge this is the first framework that integrates network definition at the granularity of empirical data with automatic visualisation and automatic model creation.

Discussion

It is clear that the complexity of signal-transduction networks is one of the major challenges in systems biology, impeding our ability to visualise, simulate and ultimately understand these networks. This issue has been widely recognised and substantial efforts have been committed to improve and standardise our tools for visualisation and modelling of

Figure 2 The reaction graph compactly displays the topology of the *S. cerevisiae* MAP kinase network. (**A**) The reaction graph of the MAPK network displays the components as nodes and the reactions as edges. Each component is defined by a central major node and peripheral minor nodes indicating domains, subdomains and specific residues (blue). When interacting domains and target residues are known, reactions are displayed as edges between these minor nodes. In contrast, the condensed reaction graph (**B**) displays each component as a single node, and each type of reaction between two nodes as a single edge. Nodes are either proteins (circles), small molecules (diamonds) or DNA (square). Edge colours indicate reaction type (co-substrates and co-products): Grey; protein–protein interaction (N/A), red; phosphorylation (− ATP, + ADP), orange; guanine nucleotide exchange (− GTP, + GDP), blue; dephosphorylation or GTPase activation (+ P_i), gold; ubiquitination (− ubiquitin, − ATP, + ADP, + P_i), black; phosphotransfer or proteolytic cleavage (N/A). The domain layout in (A) prioritises readability and domain organisation does not reflect linear sequence or protein structure. Arrowheads indicate directionality for unidirectional or reciprocal catalytic modifications. Reactions for which we found no direct evidence but which are supported by convincing genetic data has been included as dashed lines. Note the much higher frequency of reported phosphorylation reactions as compared with dephosphorylation reactions; in total the network includes 68 phosphorylation reactions but only 16 dephosphorylation reactions (A).

cellular networks (Hucka *et al*, 2003; Le Novere *et al*, 2009). These standardisation efforts are essential for data exchange and reusability, but many of the existing tools are unsuitable for definition, visualisation and mathematical modelling of large networks. The arguably most important problems are the combinatorial complexity, the granularity difference between



empirical and theoretical data, and the lack of exchange formats between different theoretical descriptions. Here, we have introduced a new framework for network definition at the same granularity as most empirical data. This format was already available for C1 (reaction) information, as our list of elemental reactions uses the same format as high-throughput data (PSICQUIC). We describe contextual information at the same granularity in our contingency list (C2), which not only allows an intuitive and accurate translation of empirical data but also largely avoids the combinatorial complexity. Contrary to state transition based descriptions but like the related rule-based format, the reaction and contingency based description becomes smaller the less knowledge we have as only known reactions and contingencies are considered. This format also provides for highly detailed referencing as each elemental reaction and contingency can and should be tied to empirical evidence (i.e., research paper(s)). Furthermore, we show that this format is stringent and unambiguously define both rule-based models and graphical formats, such as the activity flow diagram (condensed reaction graph), entity relationship diagram and process description formats of SBGN. Our framework also supports two new visualisation formats that we introduce here and that can display our complete knowledge database (the complete reaction and contingency lists). Finally, our framework provides a very high reusability and extendibility, as the underlying network definition—in list format—is very easy to extend, merge and reuse in other context, which is not the case for most graphically or mathematically defined systems. Of course, this level of definition still leaves the issues of parameter estimation and graphical layout, but these would typically need to be repeated even when merging graphical and mathematical network definitions. Hence, we advocate a more fundamental level of network definition than graphical or mathematical formalism. We envisage this or a similar framework as a standard to greatly facilitate model/network construction, exchange and reusability.

We have applied this method to map out the MAP kinase network of *S. cerevisiae*. This network was chosen as a benchmark since it is both well characterised and representative for signal transduction in general. It consists of three clear subgraphs, which have traditionally been considered more or less insulated pathways; the High Osmolarity Glycerol (Hog) pathway, the Protein Kinase C (PKC) pathway and the MATing (MAT) pathway, which almost completely overlaps with the PseudoHyphal Differentiation (PHD) pathway. These pathways have also been mapped or documented in several other efforts. KEGG presents a combined map of the traditional MAP kinase pathways in a format similar to its metabolic pathways (Kanehisa *et al*, 2006, 2010). However, the stringent edge definitions used for the metabolic networks have been abandoned and this is a 'biologist's graph'. The picture is similar with the maps of yeast MAP kinase pathways at Science STKE (e.g., Thorner *et al*, 2005). For example, these maps display Ste11 with four upstream regulators, but it is unclear how they regulate Ste11 and how their contributions combine (e.g., AND or OR?). Therefore, these network maps may provide an excellent introduction to the networks by providing a components list and a rough idea of the components' roles in the network, but they neither define reactions (C1) nor contingencies (C2) unambiguously. On the opposite end, we have the recently published process description of the cell cycle and its surrounding signalling network (Kaizu *et al*, 2010). This contains explicit definition of both C1 and C2 information. However, the tremendous number of specific states in such a network forces simplifications, which not only leads to a loss of knowledge, but also mixes up known contingencies (C2) with arbitrary assumptions made to simplify the network. One example in this particular case would be the separation of the upstream activation of Ste11 and its downstream effect on the Hog and Mating pathways. The output of this module is defined by the context of its activation, and this information is lost due to these arguably necessary simplifications. In addition, the granularity difference between the highly specific map states and the underlying biological data makes the mapping ambiguous, leading to further unsupported assumptions. Despite these shortcomings, the process description is useful for visualisation of certain network properties due to the explicit representation of highly detailed knowledge such as target residues. However, we stress that neither of these established and widely used methods are sufficient to accurately capture the entire signal-transduction network. Instead, we introduce the contingency matrix and the bipartite regulatory graph as alternative methods, which are able to fully capture the entire knowledge database without simplifications or assumptions. Together with the established methods, these visualisations provide an unprecedented view on the chosen benchmark system, and we trust that this completely referenced and comprehensive map of the MAP kinase signalling network in *S. cerevisiae* will be a useful reference material for the research community.

These results have direct bearing on the many efforts to create large data repositories. Pure reaction (C1) data, such as protein – protein interaction networks, can be retrieved using the standardised Molecular Interaction Query Language (MIQL; which our reaction list is designed to be compatible with) and PSICQUIC (PSICQUIC). PSICQUIC accesses, for example, ChEMBL (Overington, 2009), BioGrid (Breitkreutz *et al*, 2010), IntAct (Aranda *et al*, 2010), DIP (Xenarios *et al*, 2002), MatrixDB (Chautard *et al*, 2009) and Reactome (Croft *et al*, 2010). Several of these databases have additional information including contingency (C2) information and a standardised (non-graphical) format for definition and

Figure 3 The condensed reaction graph is an excellent tool for visualisation of high-throughput data. (**A**) Physical interactions within the MAPK network. The global protein–protein interaction network was retrieved from Biogrid (Stark *et al*, 2006), filtered for physical interactions excluding two hybrid, and visualised on the condensed reaction graph (Figure 2A). Purple edges indicate protein–protein interactions and their thickness indicates the number of times they were picked up, ranging from a single time (dashed line) to 19 times. Nodes that appear faded have no interactions with any other component in the MAPK network reported in this data set. Note that the nodes that do not correspond to single ORFs would be excluded automatically (e.g., the SCF complex, DNA, lipids). The smaller, boxed network display the corresponding two-hybrid interaction network. (**B**) Genetic interactions within the MAPK network. Synthetic lethal interactions were retrieved from Biogrid and visualised as per (A). Also quantitative data, such as mutant phenotypes and gene expression levels, can be directly visualised on the network.

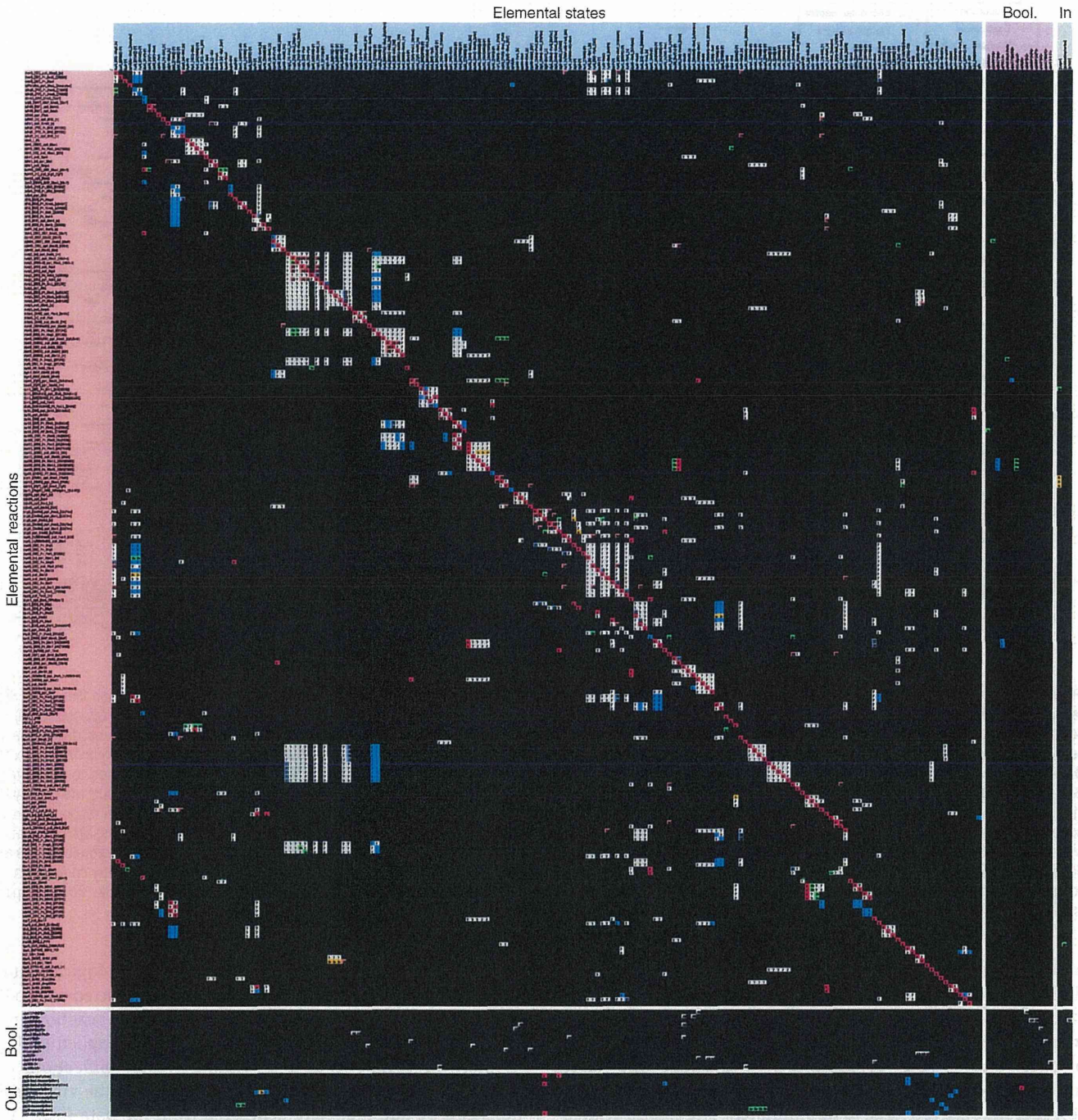


Figure 4 The contingency matrix provides a complete description of the network or network module. The core contingency matrix is spanned by the elemental reactions (rows, in red) and the elemental states (columns, in blue). The additional blocks are derived from the contingency list and contain the formation rules (rows) and effects (columns) of Boolean states (both purple) as well as the output of (rows) and input to (columns) the network (both grey). The cells in the matrix define how each reaction (row) depends on each state (column). The effects range from being absolutely required ('!'), via positive effector ('K+'), no effect ('0') and negative effector ('K–') to absolutely inhibitory ('x'), or it can be unknown or undefined ('?'). Each Boolean state is defined by a single operator ('AND' or 'OR') for the elemental states, other Booleans and/or inputs that defines it. The contingency matrix displayed here contains the complete MAPK network. Note that the contingency matrix is sparsely populated. This is both because most combinations of reactions and states lack overlap in components (black squares) and because we have very limited knowledge of the possible contingencies (grey squares). Overall, the information on what reactions can occur is much more abundant than on how they are regulated.

retrieval would further improve the usefulness of these resources and facilitate further analysis of the stored information. The framework we propose here provides such a format with the key advantage of including export to mathematical models. Since mathematical modelling is the most central and natural step to bring the knowledge in these databases into a useful form, where quantitative systems properties can most exhaustively be analysed, the introduction of such an export is

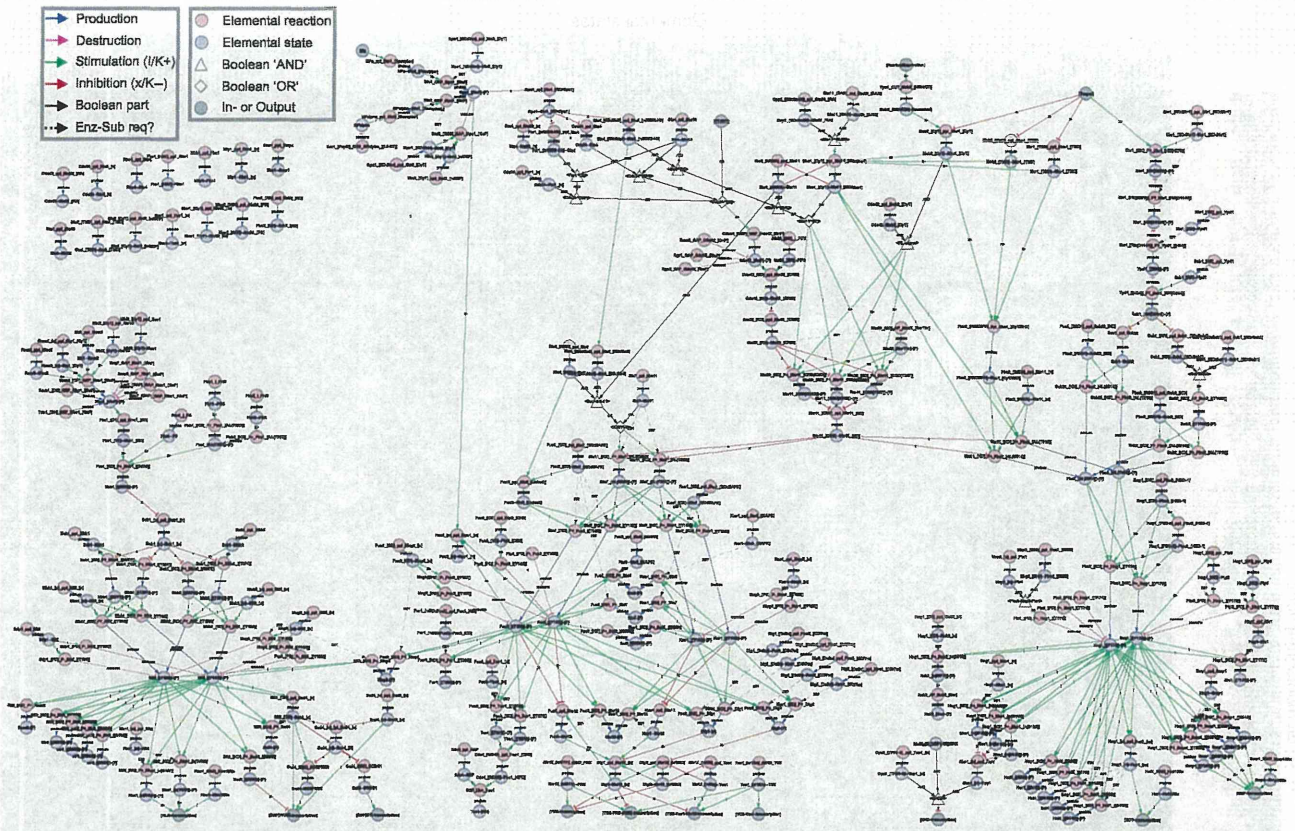


Figure 5 The regulatory graph visualise the causality between reactions and reveals the regulatory structure of the network. This bipartite graph illustrates the relationships between the reactions (red nodes) and states (blue nodes) within the network. Edges from reactions to states define how states are produced (blue) or consumed (purple), and each such edge corresponds to a single elemental reaction. Edges from states to reactions define how states regulate other reactions, and each such edge correspond to a single contingency (Green; absolute requirement ('!') or positive effector ('K + '), red; negative effector ('K–') or absolutely inhibitory ('x')). Booleans are used when the effect on a reaction cannot be attributed to single elemental states (white diamonds (OR) or triangles (AND) connected to the states/Booleans/inputs that define them with black lines). Inputs are displayed in grey and connected to the elemental reaction(s) they influence. Likewise, outputs are displayed in grey and connected to the states they are influenced by. Signals can be followed through the network from external cues (grey; top) to transcriptional response (grey; bottom) as all edges are directional. Reactions without input are not (known to be) regulated and would therefore be expected to have constant rates; likewise states without output have no (defined) impact on the system. We have also included likely but undocumented requirements for enzyme–substrate bindings before catalysis as dashed lines. The regulatory graph is the only graphical representation using the complete information in the contingency matrix, and hence the only complete and completely graphical visualisation of the network. It is also the most potent visualisation to evaluate the degree of knowledge about the network. For example, visualisation of high-throughput data would result in disconnected reaction–state pairs only, due to the lack of regulatory information (no C2 data).

an important step forward. This framework is still not as flexible as direct model definition but it provides distinct advantages. Formulating models directly using classical state transition reactions is either subjective or very cumbersome in practice due to the combinatorial explosion, and state transition based models for the networks of the size we consider here are too large to be simulated. The closest related modelling framework is rule-based modelling, in which models can be formulated without these combinatorial explosion problems, and it is also to a rule-based format that we export our models. However, the classical rule-based modelling frameworks lack all the database properties of our framework, such as the contingency matrix and its export to various novel visualisation formats. In short, one could therefore say that our framework combines the best of existing knowledge databases with new visualisation tools and rule-based modelling.

In conclusion, we present a method to document and visualise signal-transduction networks that improves on previous strategies in the following respects; (I) it allows concise mapping at the same granularity as biological data, hence pre-empting the need for implicit, unsupported assumptions, (II) it allows referencing of each elemental reaction and contingency separately and handles unknowns explicitly, (III) the network can be visualised without any simplifications or assumptions that increase the uncertainty, (IV) the visualisations can be automatically generated from the data files, (V) the network definition is a template from which a mathematical model can be automatically generated (VI) and exported to SBML and (VII) the supplied template and *rxncon* tool makes the method immediately useful for anyone with an interest in signal transduction. Hence, our framework bridge three critical levels of signal-transduction network analysis; definition, visualisation and mathematical modelling, as well as empirical data and theoretical analysis.

Materials and methods

The MAP kinase network map is based on the papers listed below. The specific reference(s) are listed for each reaction and contingency