Giovanni Boniolo
FIRC Institute of Molecular Oncology (IFOM)- Milano
European School of Molecular Medicine (SEMM) - Milano
Faculty of Medicine - University of Milano

Could We Formalize Biochemical Processes? And why?

- The theoretical framework
- The proposal
- The possible research programme

mathematics	geometry	logic	dynamics (differential equations)

Classical mechanics	Vectorial calculus	Euclidean geometry	Classical logic	Newton law
General relativity	Differential topology	Riemannian geometry	Classical logic	Einstein equations
Quantum mechanics	Hilbert space	Euclidean geometry	Quantum logic	Schrödinger equation

Molecular biology	NO	NO	**YES**	YES

From statements to molecules

From classical logic to resource logic

NON CLASSICAL COMPUTATIONAL LOGIC

ZSYNTAX

	Linguistic case	**Biological case**
Premises (Hypothesis): Γ	Conjunction of statements	Aggregate of molecules
Inferential process	Classical logic	Biochemical reactions (non classical logic)
Conclusion (Thesis): C	Statement	Aggregate of molecules

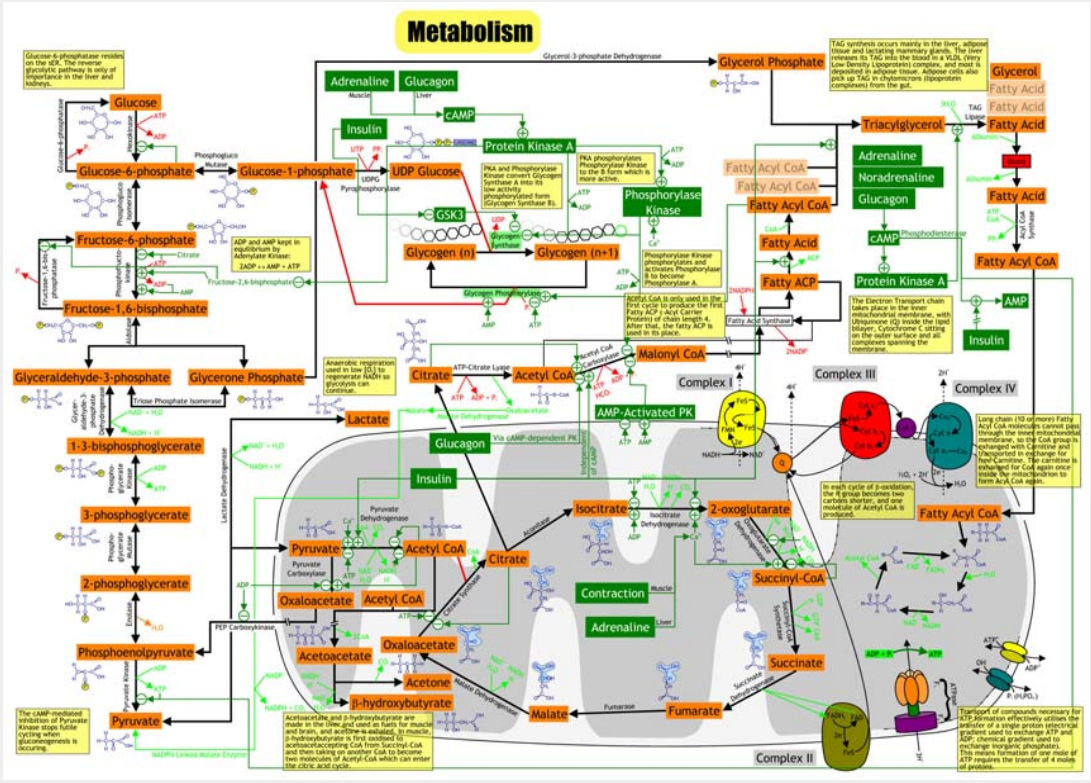
$$\Gamma \mid\!\!- \; \mathbf{C}$$

2 EXAMPLES

Γ= (D-glucose, hexokinase, phosphoglucoisomerase, phosphofructokinase, ATP, ATP)

Theorem 1:

Γ |— (fructose-1,6-phosphate)



Theorem 2:

(*MDM2*, p53) |— MDM2

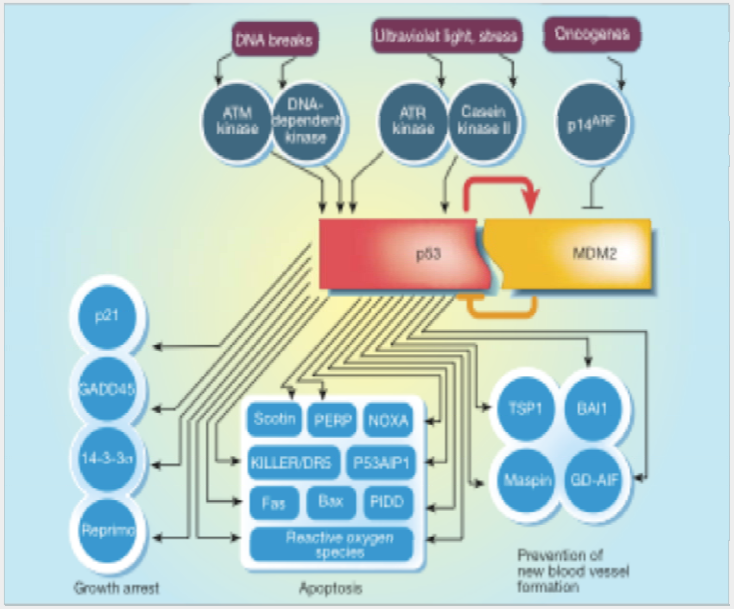
Nature **408** 307 (2000)

news and views feature

Surfing the p53 network

Bert Vogelstein, David Lane and Arnold J. Levine

The p53 tumour-suppressor gene integrates numerous signals that control cell life and death. As when a highly connected node in the Internet breaks down, the disruption of p53 has severe consequences.



The three operators of our ZSYNTAX

1) Z-interaction (indicated by ⊙)

we denote by A⊙B the type of molecules which results from the interaction of two types of molecules A and B.

p53⊙*MDM2*

2) Z-conjunction (indicated by &)

we denote by A&B&C&D the type of aggregate constitutes by the types of molecules A, B, C, D.

(D-glucose & hexokinase & phosphoglucoisomerase & phosphofructokinase)

3) Z-conditional (indicated by →)

we denote by A&C →B the fact that there is a transition path from an aggregate of type A&C to an aggregate of type B.

(p53 & *MDM2*_) → (p53⊙*MDM2*).

The two kinds of valid formulas for ZSYNTAX

1) The **empirically valid formulas** (EVF). They represent singular reactions and their validity depends on the fact that the processes they describe are empirically corroborated. We can have EVFs representing

(i) that two molecules interact, e.g.

p53 & *MDM2*_ → p53⊙*MDM2*;

(ii) that the interaction delivers certain products in a biochemical sense, e.g.

D-glucose-6-phosphate ⊙ phosphoglucoisomerase → D-fructose-6-phosphate)

(iii) that the interaction allows the gene expression, e.g.,

p53⊙*MDM2* → MDM2.

2) The **logically valid formulas**. Their validity rests only on the definitions of the logical operators inside the language we have constructed. They give the rules to move from one EVF to another EVF

Example of a logically valid formula:

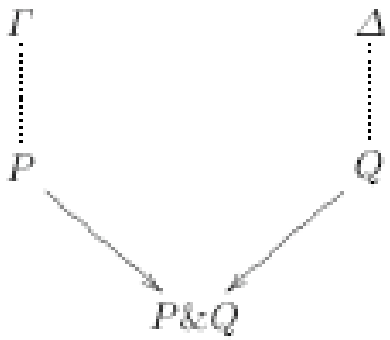
$(A \rightarrow B \ \& \ A) \rightarrow B$

This can be expressed as the *modus ponens rule:*

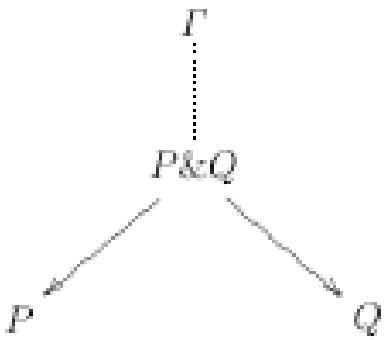
1. If A then B	1. A $\rightarrow$B	1. (p53 & *MDM2*) $\rightarrow$ (p53$\odot$ *MDM2*)
2. A	2. A	2. (p53 & *MDM2*)
B	B	(p53$\odot$ *MDM2*)

Our primitive logical rules for & are the following:

&-introduction

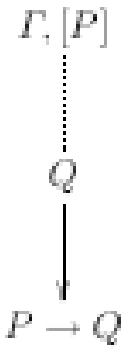


&-elimination

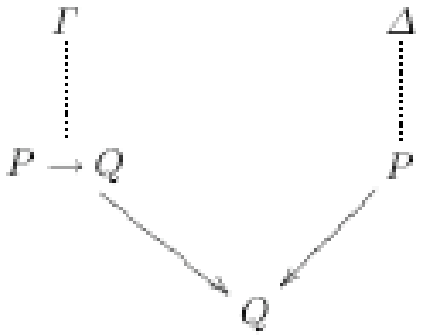


The rules for $\rightarrow$ are the following:

$\rightarrow$-introduction



$\rightarrow$-elimination

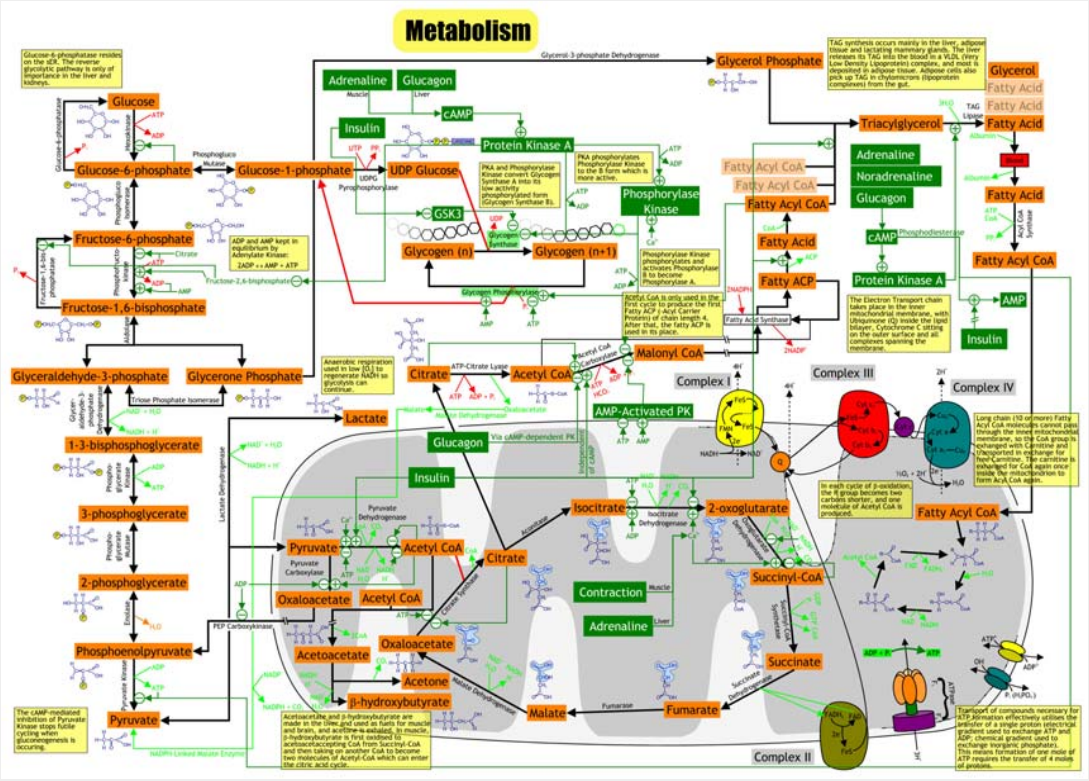


2 EXAMPLES OF DEMONSTRATION

Γ= (D-glucose, hexokinase, phosphoglucoisomerase, phosphofructokinase, ATP, ATP)

Theorem 1:

Γ |— (fructose-1,6-phosphate)



1. Glc & HK & GPI & PFK & ATP & ATP	IA
2. Glc & HK	From 1 by &*E*
3. GPI	From 1 by &*E*
4. PFK	From 1 by &*E*
5. ATP	From 1 by &*E*
6. ATP	From 1 by &*E*
7. Glc & HK → Glc⊙HK	EVF
8. Glc⊙HK	From 2,7 by →*E*
9. (Glc⊙HK) & ATP	From 5,8 by &*I*
10. (Glc⊙HK) & ATP → (Glc⊙HK)⊙ATP	EVF
11. (Glc⊙HK)⊙ATP	From 9,10 by →*E*
12. (Glc⊙HK)⊙ATP → G6P & HK & ADP	EVF
13. G6P & HK & ADP	From 11,12 by →*E*
14. G6P	From 13 by &*E*
15. HK	From 13 by &*E*
16. ADP	From 13 by &*E*
17. G6P & GPI	From 3,14 by &*I*
18. G6P & GPI → G6P⊙GPI	EVF
19. G6P⊙GPI	From 17,18 by →*E*
20. G6P⊙GPI → F6P & GPI	EVF
21. F6P & GPI	From 19,20 by →*E*
22. F6P	From 21 by &*E*
23. GPI	From 21 by &*E*
24. F6P & PFK	From 4,22 by &*I*
25. F6P & PFK → F6P⊙PFK	EVF
26. F6P⊙PFK	From 24,25 by →*E*
27. (F6P⊙PFK) & ATP	From 6,26 by &*I*
28. (F6P⊙PFK) & ATP → (F6P⊙PFK)⊙ATP	EVF
29. (F6P⊙PFK)⊙ATP	From 27,28 by →*E*
30. (F6P⊙PFK)⊙ATP → F1,6P & PFK & ADP	EVF
31. F1,6P & PFK & ADP	From 29,30 by →*E*
32. F1,6P	From 31 by &*E*
33. (Glc & HK & GPI & PFK & ATP & ATP) → F1,6P	From 1-32 by →*I*

Theorem

Glc&HK&GPI&PFK&ATP&ATP ⊢ F1,6P

Demonstration

1. Glc & HK → Glc⊙HK
2. (Glc⊙HK) & ATP → (Glc⊙HK)⊙ATP
3. (Glc⊙HK)⊙ATP → G6P & HK & ADP
4. G6P & GPI → G6P⊙GPI
5. G6P⊙GPI → F6P & GPI
6. F6P & PFK → F6P⊙PFK
7. (F6P⊙PFK) & ATP → (F6P⊙PFK)⊙ATP
8. (F6P⊙PFK)⊙ATP → F1,6P & PFK & ADP

Theorem 2:

(*MDM2*, p53) |— MDM2

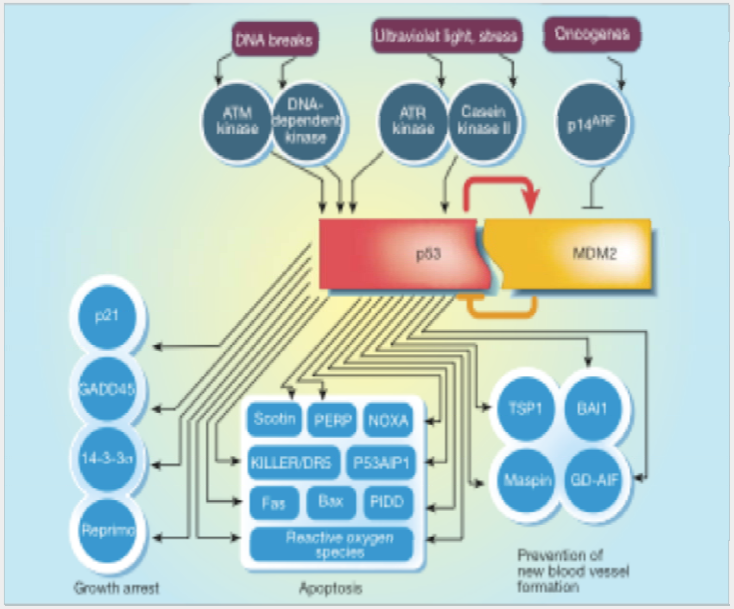
Nature **408** 307 (2000)

news and views feature

Surfing the p53 network

Bert Vogelstein, David Lane and Arnold J. Levine

The p53 tumour-suppressor gene integrates numerous signals that control cell life and death. As when a highly connected node in the Internet breaks down, the disruption of p53 has severe consequences.



1.	TP53 & TP53 & *MDM2* & U & P	IA
2.	TP53	From 1 by &*E*
3.	TP53	From 1 by &*E*
4.	*MDM2*	From 1 by &*E*
5.	U	From 1 by &*E*
6.	P	From 1 by &*E*
7.	TP53 & *MDM2*	From 2,4 by &*I*
8.	TP53 & *MDM2* → TP53⊙*MDM2*	EVF
9.	TP53⊙*MDM2*	From 7,8 by →*E*
10.	TP53⊙*MDM2* → MDM2	EVF
11.	MDM2	From 9,10 by →*E*
12.	MDM2 & TP53	From 3,11 by &*I*
13.	MDM2 & TP53 → MDM2⊙TP53	EVF
14.	MDM2⊙TP53	From 12,13 by →*E*
15.	(MDM2⊙TP53) & U	From 5,14 by &*I*
16.	(MDM2⊙TP53) & U → (MDM2⊙TP53)⊙U	EVF
17.	(MDM2⊙TP53)⊙U	From 15,16 by →*E*
18.	(MDM2⊙TP53)⊙U → MDM2 & (TP53⊙U)	EVF
19.	MDM2 & (TP53⊙U)	From 17,18 by →*E*
20.	TP53⊙U	From 19 by &*E*
21.	(TP53⊙U) & P	From 6,20 by &*I*
22.	(TP53⊙U) & P → (TP53⊙U)⊙P	EVF
23.	(TP53⊙U)⊙P	From 21,22 by →*E*
24.	(TP53⊙U)⊙P → d(TP53) & U & P	EVF
25.	d(TP53) & U & P	From 23,24 by →*E*
26.	d(TP53)	From 25 by &*E*
	(TP53 & TP53 & *MDM2* & U & P) → d(TP53)	From 1-26 by →*I*

Theorem

TP53 & TP53 & *MDM2* & U & P ⊢ d(TP53)

Demonstration

1. TP53 & *MDM2* → TP53⊙*MDM2*
2. TP53 ⊙ *MDM2* → MDM2
3. MDM2 & TP53 → MDM2⊙TP53
4. (MDM2 ⊙TP53) & U → (MDM2⊙TP53)⊙U
5. (MDM2⊙TP53)⊙U → MDM2 & (TP53⊙U)
6. (TP53⊙U) & P → (TP53⊙U)⊙P
7. (TP53⊙U)⊙P → d(TP53) & U & P

Why should we use it?

1) To formalize molecular biology

2) To perform text mining

3) To predict biological reactions and biological products

TEXT MINING

Seeking a New Biology through Text Mining

Andrey Rzhetsky,[1,*] Michael Seringhaus,[2] and Mark Gerstein[2]
[1]University of Chicago, Chicago, IL 60637, USA
[2]Yale University, New Haven, CT 06510, USA
*Correspondence: arzhetsky@uchicago.edu
DOI 10.1016/j.cell.2008.06.029

Cell 134, July 11, 2008 ©2008 Elsevier Inc.

Tens of thousands of biomedical journals exist, and the deluge of new articles in the biomedical sciences is leading to information overload. Hence, there is much interest in text mining, the use of computational tools to enhance the human ability to parse and understand complex text.

Imagine that a graduate student enters the U.S. Library of Congress with the goal of retrieving all texts relevant to protein glycosylation. Her problem is straightforward, known among text miners as *information retrieval* (IR). If the student must not only find the books but also flag the most important concepts she encounters in each, she is performing *named entity recognition* (NER). Undaunted by her workload, imagine she decides to identify relations between concepts, such as "protein BAD binds to protein BAX" (called *information extraction* or IE). Then she takes on additional tasks such as question/answer (QA) and text summarization (TS). Computational IR, NER, IE, QA, and TS are all part of text mining and belong to the larger field of *natural language processing* (NLP), which itself is a part of *artificial intelligence* (AI) that aims to recreate or surpass the computational ability of the human brain. Although multiple definitions exist, text mining is typically associated with information retrieval, extraction, and synthesis, with a special emphasis on gaining new knowledge (Table 1).

available online). This is because biology and medicine are unusually rich in terminology: the collective vocabulary used by biomedicine incorporates many millions of terms. The exact number is unknown and constantly in flux. Because this vocabulary is large and dynamic, new terms emerge rapidly and erratically. As a result, the same real-world object may have numerous names (synonyms), whereas distinct objects can be identified with the same name (homonyms). The terms that most notoriously suffer from synonym and homonym abundance are gene and protein names (Hirschman et al., 2002; Wilbur et al., 1999). A given gene may be denoted by several dozen synonymous names—for example, the *Drosophila* genes *br* and *mod(mdg4)* have 82 and 64 aliases, respectively. Even worse, vastly different and incompatible naming systems are used in different species, and gene aliases themselves merge into intricate semantic networks that connect gene names, pop-culture catch phrases, idioms, and common everyday utterances borrowed from multiple languages.

PREDICTION

Abduction

Let us suppose that we know the final molecule of a biological process (call it B), but we do not know exactly

either **1)** which is the initial aggregate (call it Γ),

or **2)** which are the right biological reactions leading from that initial aggregate to the final molecule.

In this case the automated prover may indicate that something is missing:

maybe **1)** an additional molecule in the initial aggregate (that is, we should find a new molecule),

maybe **2)** an additional reaction (that is, we should find a new reaction).

But not only A well-constructed automated prover should be able to extract a range of alternative hypotheses concerning how to adjust the data and/or the reactions involved and, thus, to indicate us what should miss.

At this point, we could go back to the lab and investigate the hypotheses which have been automatically generated by the computer.

Typically, in automated theorem proving, one starts from the conclusion B and looks back for the possible premises Γ.

Reiterating this inverse process the automated prover constructs several possible paths. Going through all possible paths (and using suitable heuristics to prune the wrong ones), it eventually arrives at the right initial aggregate of molecules Γ.

"Human" prover	Automated prover
Γ ⋮ B	Γ ? ? ? ? ? B

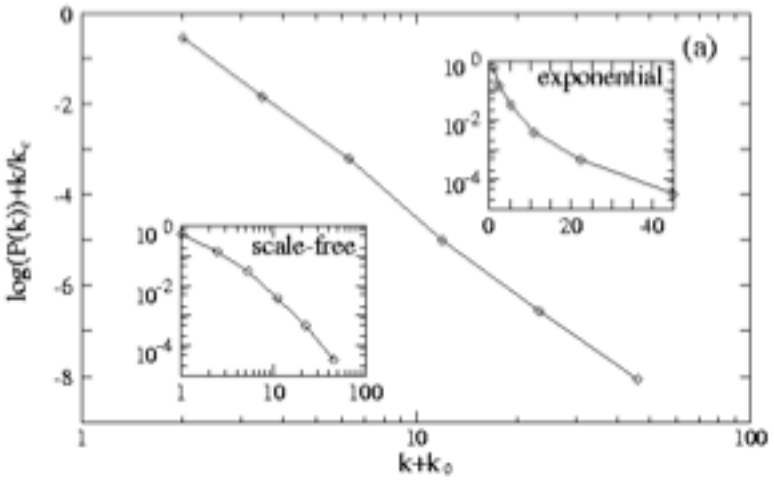
Are there competitors to ZSYNTAX?

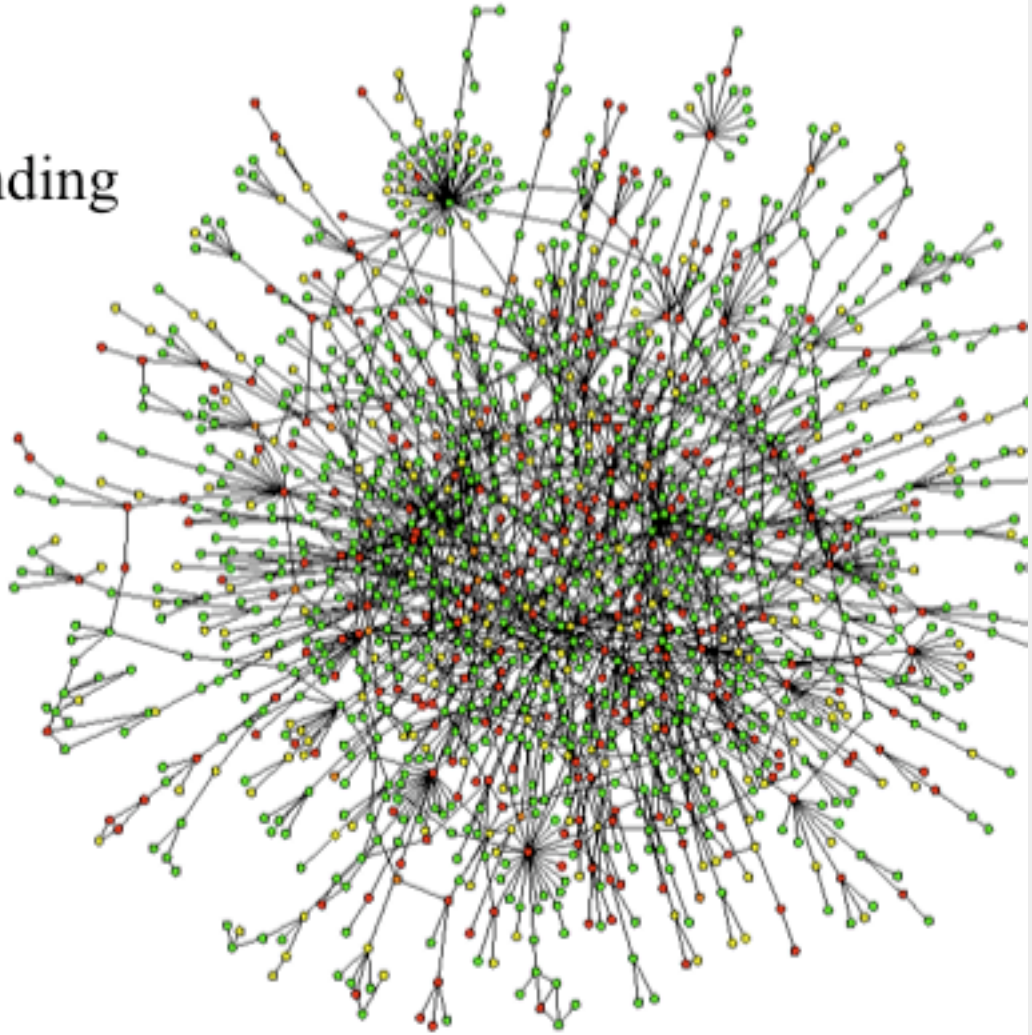
Zsyntax and biological networks

Topology of the protein network

Nodes: proteins

Links: physical interactions-binding





$$P(k) \sim (k+k_0)^{-\gamma} \exp(-\frac{k+k_0}{k_\tau})$$

H. Jeong, S.P. Mason, A.-L. Barabasi, Z.N. Oltvai, Nature 411, 41-42 (2001)

To elucidate the age of retrogene movements, we dated the human duplications involving X-linked parents or retrogenes both by comparison to the mouse genome sequence and by sequence divergence analysis (*16*). Most copies that escape X linkage (12/15) as well as most copies that obtain X linkage (10/13) originated before the human-mouse split (Fig. 2, tables S7 and S8). Duplicates in the mouse genome show the same pattern, consistent with this notion. Thus, both patterns result from ancient evolutionary forces common to eutherian mammals. However, this process appears to be an ongoing characteristic of eutherian X evolution, because 6/28 events have occurred subsequent to the human-mouse split in the human lineage, 6/33 retropositions have occurred within the past ~80 million years in the mouse lineage, and some of these retroduplicate pairs have high sequence similarity (>95%) at synonymous sites. This chromosome-biased gene origination appears to be an important process actively driving the differentiation of the X chromosome in mammals and suggests that this differentiation is still in progress.

References and Notes

1. B. T. Lahn, N. M. Pearson, K. Jegalian, *Nature Rev. Genet.* **2**, 207 (2001).
2. H. Skaletsky *et al.*, *Nature* **423**, 825 (2003).
3. J. A. Marshall Graves *et al.*, *Cytogenet. Genome Res.* **96**, 161 (2002).
4. J. R. McCarrey, *Bioscience* **44**, 20 (1994).
5. M. J. Larcher, A. O. Urrutia, L. D. Hurst, *Mol. Biol. Evol.* **20**, 1113 (2003).
6. P. J. Wang, J. R. McCarrey, F. Yang, D. C. Page, *Nature Genet.* **27**, 422 (2001).
7. J. C. Venter *et al.*, *Science* **291**, 1304 (2001).
8. B. Lewin, *Genes VII* (Oxford University Press, New York, 2000).
9. E. Betran, K. Thornton, M. Long, *Genome Res.* **12**, 1854 (2002).
10. P. M. Harrison *et al.*, *Genome Res.* **12**, 272 (2002).
11. L. Z. Strichman-Almashanu, M. Bustin, D. Landsman, *Genome Res.* **13**, 800 (2003).
12. E. Betrán, W. Wang, L. Jin, M. Long, *Mol. Biol. Evol.* **19**, 654 (2002).
13. J. Brosius, *Science* **251**, 753 (1991).
14. E. S. Lander *et al.*, *Nature* **409**, 860 (2001).
15. T. Hubbard *et al.*, *Nucleic. Acids. Res.* **30**, 38 (2002).
16. Materials and methods are available as supporting material on *Science* Online.
17. R. H. Waterston *et al.*, *Nature* **420**, 520 (2002).
18. Z. Zhang, P. Harrison, M. Gerstein, *Genome Res.* **12**, 1466 (2002).
19. C.-I. Wu, E. Y. Xu, *Trends Genet.* **19**, 243 (2003).

A Map of the Interactome Network of the Metazoan *C. elegans*

Siming Li,[1]† Christopher M. Armstrong,[1]* Nicolas Bertin,[1]† Hui Ge,[1]* Stuart Milstein,[1]† Mike Boxem,[1]† Pierre-Olivier Vidalain,[1]* Jing-Dong J. Han,[1]† Alban Chesneau,[1,2]* Tong Hao,[1] Debra S. Goldberg,[3] Ning Li,[1] Monica Martinez,[1] Jean-François Rual,[1,4] Philippe Lamesch,[1,4] Lai Xu,[5]† Muneesh Tewari,[1] Sharyl L. Wong,[3] Lan V. Zhang,[3] Gabriel F. Berriz,[3] Laurent Jacotot,[1]‡ Philippe Vaglio,[1]‡ Jérôme Reboul,[1]§ Tomoko Hirozane-Kishikawa,[1] Qianru Li,[1] Harrison W. Gabel,[1] Ahmed Elewa,[1]‖ Bridget Baumgartner,[5] Debra J. Rose,[6] Haiyuan Yu,[7] Stephanie Bosak,[8] Reynaldo Sequerra,[8] Andrew Fraser,[9] Susan E. Mango,[10] William M. Saxton,[6] Susan Strome,[6] Sander van den Heuvel,[11] Fabio Piano,[12] Jean Vandenhaute,[4] Claude Sardet,[2] Mark Gerstein,[7] Lynn Doucette-Stamm,[8] Kristin C. Gunsalus,[12] J. Wade Harper,[5]† Michael E. Cusick,[1] Frederick P. Roth,[3] David E. Hill,[1]¶ Marc Vidal[1]¶#

To initiate studies on how protein-protein interaction (or "interactome") networks relate to multicellular functions, we have mapped a large fraction of the *Caenorhabditis elegans* interactome network. Starting with a subset of metazoan-specific proteins, more than 4000 interactions were identified from high-throughput, yeast two-hybrid (HT–Y2H) screens. Independent coaffinity purification assays experimentally validated the overall quality of this Y2H data set. Together with already described Y2H interactions and interologs predicted *in silico*, the current version of the Worm Interactome (WI5) map contains ~5500 interactions. Topological and biological features of this interactome network, as well as its integration with phenome and transcriptome data sets, lead to numerous biological hypotheses.

To further understand biological processes, it is important to consider protein functions in the context of complex molecular networks. The study of such networks requires the availability of proteome-wide protein-protein interaction, or "interactome," maps. The yeast *Saccharomyces cerevisiae* has been used to develop a eukaryotic unicellular interactome map (*1*–*6*). *Caenorhabditis elegans* is an ideal model for studying how protein networks relate to multicellularity. Here we investigate its interactome network with HT-Y2H.

As Y2H baits, we selected a set of 3024 worm predicted proteins that relate directly or indirectly to multicellular functions (*7*). Gateway-cloned open reading frames (ORFs) were available in the *C. elegans* ORFeome 1.1 (*8*) for 1978 of these selected proteins. Of these, 81 autoactivated the Y2H *GAL1::HIS3* reporter gene as Gal4 DNA binding domain fusions (DB-X), and 24 others conferred toxicity to yeast cells. The remaining 1873 baits were screened against two different Gal4 activation domain libraries (AD-wrmcDNA and

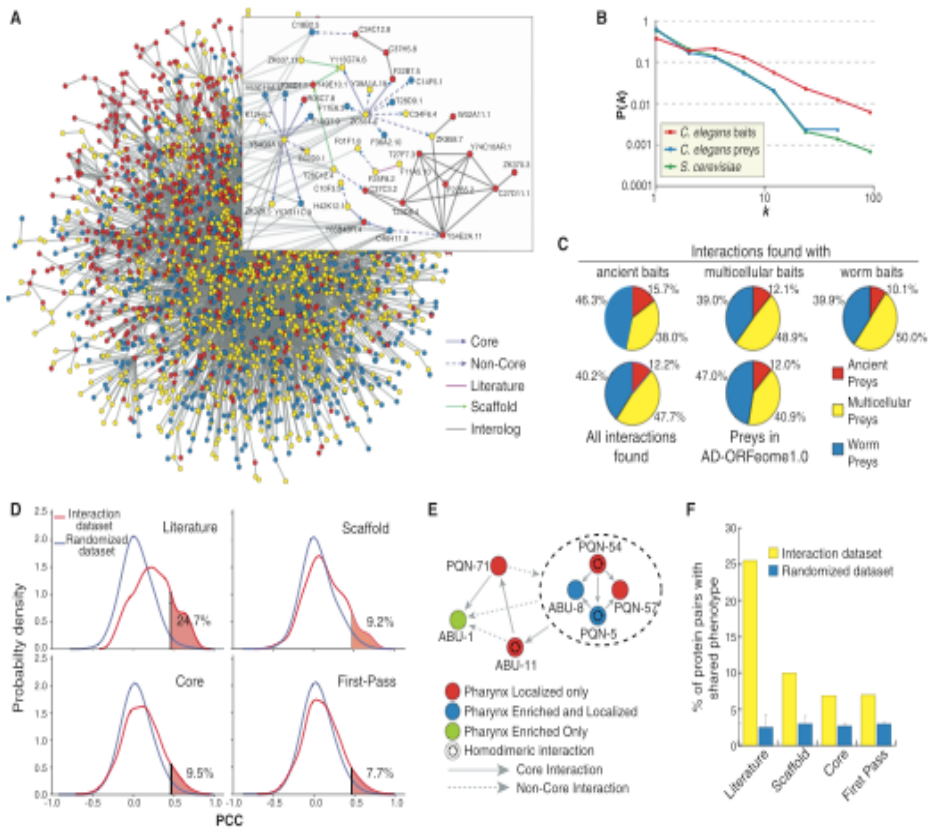


Fig. 2. Analysis of the WI5 network. (**A**) Nodes (representing proteins) are colored according to their phylogenic class: ancient (red), multicellular (yellow), and worm (blue). Edges represent protein-protein interactions. The inset highlights a small part of the network. (**B**) The proportion of proteins, $P(k)$, with different numbers of interacting partners, k, is shown for *C. elegans* proteins used as baits or preys and for *S. cerevisiae* proteins. (**C**) The pie charts show the proportion of interacting preys found in Y2H screens that fall into each phylogenic class. Also shown is the distribution of all preys found and all preys searched in the AD-ORFeome1.0 library. (**D**) Overlap with transcriptome (see text) (*18*), Pearson correlation coefficients (PCCs) were calculated and graphed for each pair of proteins in the interaction data sets and their corresponding randomized data sets. The red area to the right corresponds to interactions that show a significant relationship to expression profiling data ($P < 0.05$). (**E**) Interactions between proteins in Topomap mountain 29 (*18*). The dash-circled proteins belong to the same paralogous family (sharing more than 80% homology) and are thus collapsed into one set of interactions. (**F**) Proportion of interaction pairs where both genes are embryonic lethal ($P < 10^{-7}$).

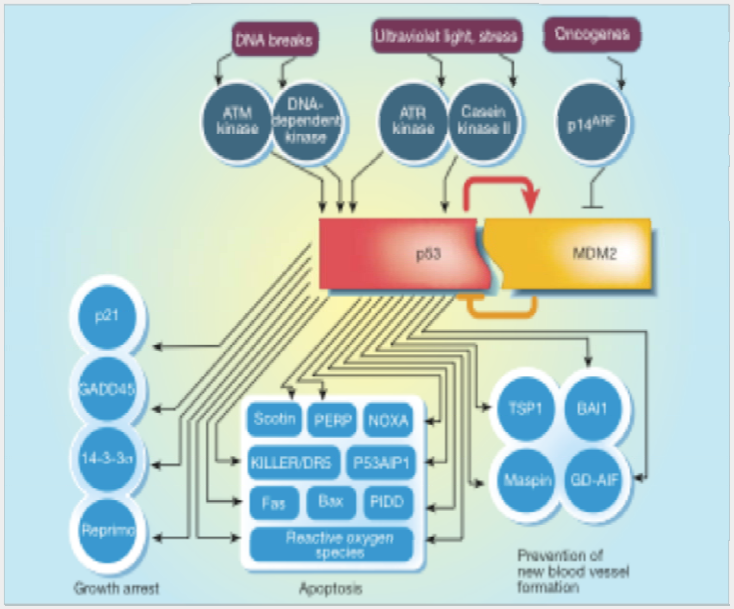
Nature **408** 307 (2000)

news and views feature

Surfing the p53 network

Bert Vogelstein, David Lane and Arnold J. Levine

The p53 tumour-suppressor gene integrates numerous signals that control cell life and death. As when a highly connected node in the Internet breaks down, the disruption of p53 has severe consequences.



Molecular Biology of the Cell
Vol. 10, 2703–2734, August 1999

Molecular Interaction Map of the Mammalian Cell Cycle Control and DNA Repair Systems

Kurt W. Kohn*

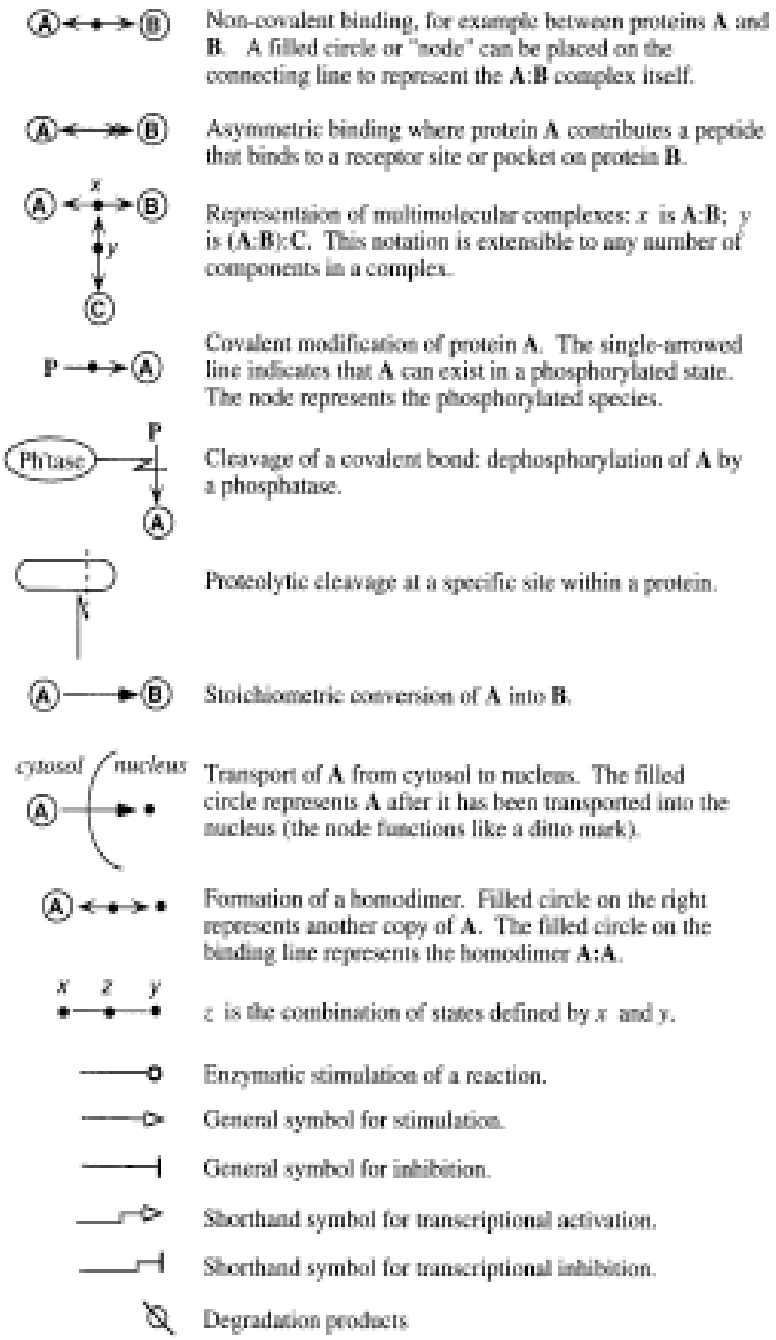
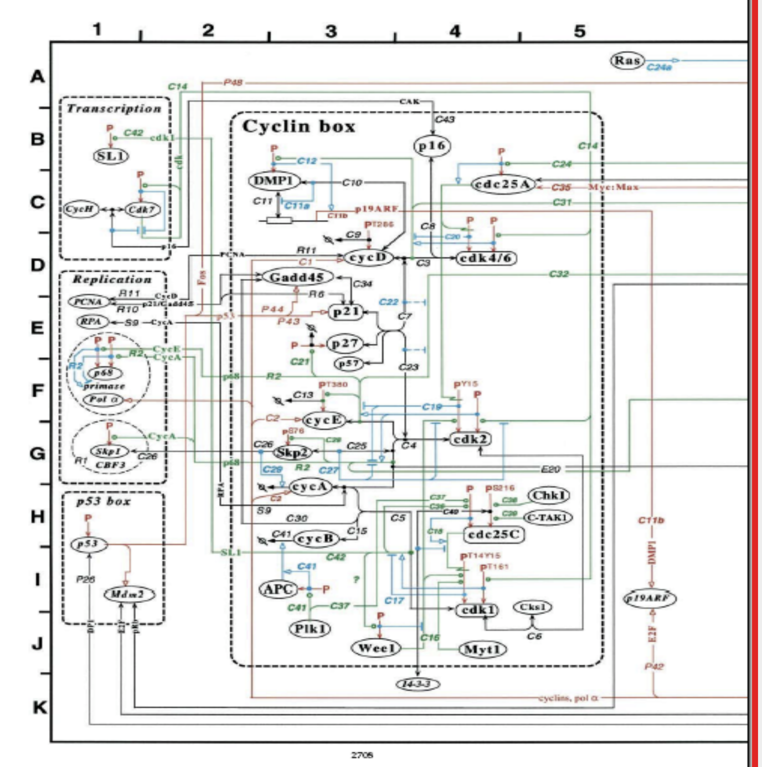
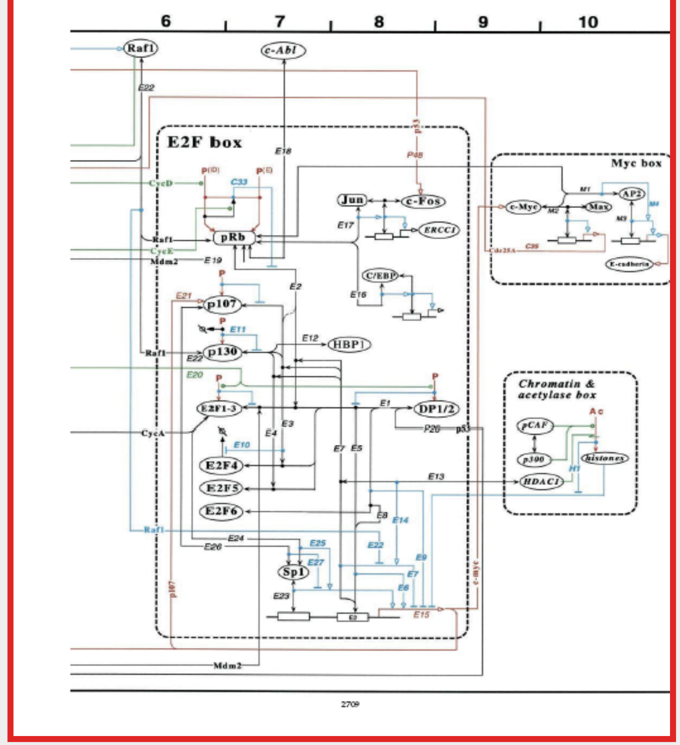
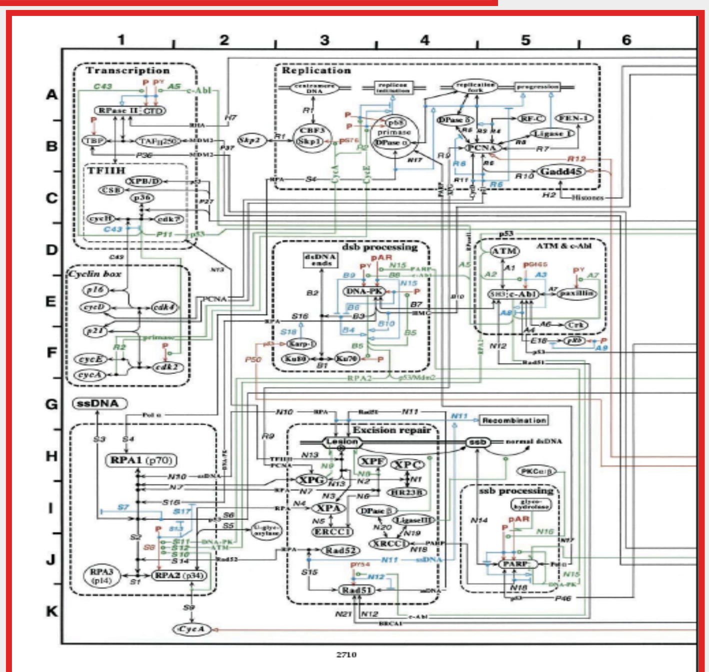
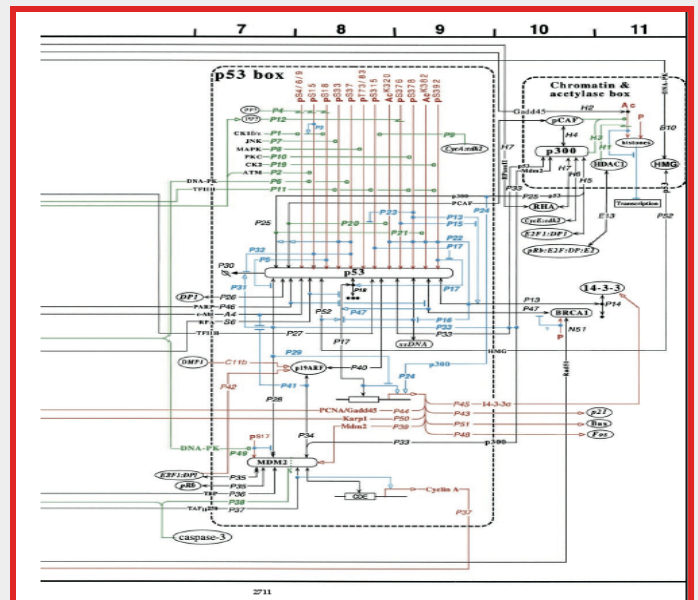


Figure 1. Summary of symbols.


Transcription
Cyclin box
Replication
p53 box
SL1
CycH
Cdk7
DMP1
cdc25A
p16
cycD
cdk4/6
Gadd45
p21
p27
p57
cycE
cdk2
Skp2
cycA
cycB
cdc25C
Chk1
C-TAK1
APC
Plk1
cdk1
Cks1
Wee1
Myt1
14-3-3
p19ARF
Ras
PCNA
RPA
p68
primase
Pol α
Skp1
CBF3
p53
Mdm2
2708

Raf1
c-Abl
E2F box
Myc box
pRb
Jun
c-Fos
ERCC1
C/EBP
p107
p130
HBP1
E2F1-3
DP1/2
E2F4
E2F5
E2F6
Sp1
c-Myc
Max
AP2
E-cadherin
Chromatin & acetylase box
pCAF
p300
histones
HDAC1
2709

Transcription
TFIIH
Replication
Cyclin box
dsb processing
ATM & c-Abl
ssDNA
Excision repair
ssb processing
Recombination
RPA1 (p70)
RPA2 (p34)
RPA3 (p14)
DNA-PK
Ku80
Ku70
ATM
c-Abl
paxillin
Crk
XPG
XPF
XPC
XPA
ERCC1
HR23B
Rad52
Rad51
XRCC1
PARP
Gadd45
PCNA
Ligase I
FEN-1
RF-C
CycA
2710

p53 box
Chromatin & acetylase box
pCAF
p300
HDAC1
HMG
p53
MDM2
p19ARF
DMP1
BRCA1
14-3-3
caspase-3
p21
Bax
Fos
2711

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, Shinagawa, 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)

Published online 4 August 2005; doi:10.1038/nbt1111

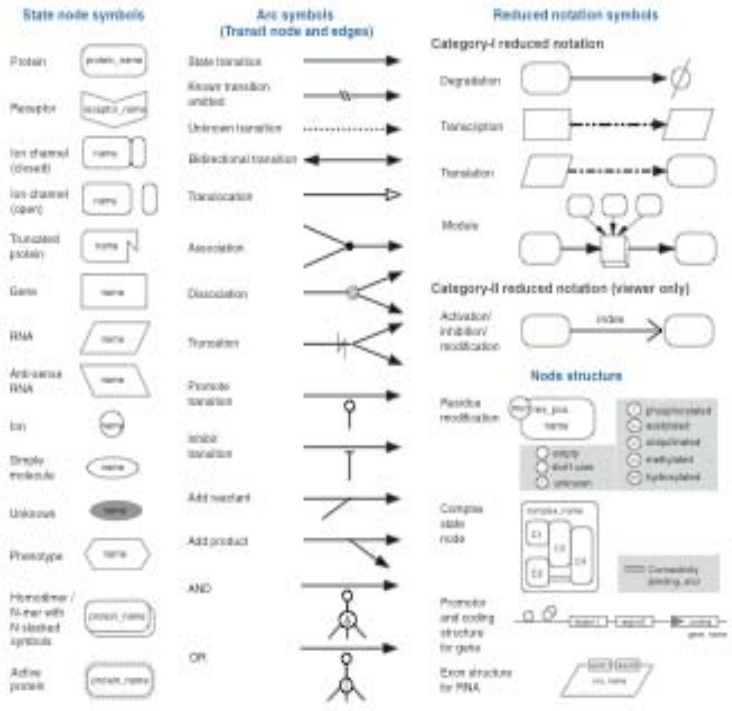


Figure 2 Proposed set of symbols for representing biological networks with process diagrams. Symbols in the process diagrams consist of visual icons for state nodes and arcs. Each arc consists of a transit node and edges. Currently, there are four reduced notations that display simplified diagrammatic 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.

Theorem

Nec. Cond. (the unpacking side)

Given a biological network satisfying the condition C^1, …, C^n, it can be rewritten as a conjunction of N Zsyntax theorems

Suff. Cond. (the packing side)

Given a conjunction of N Zsyntax theorems satisfying the condition R^1, …, R^m, it can be rewritten as a biological network

Is ZSYNTAX too simple

to grasp biological complexity?

- *Informational content and context dependence*
- *Function and context dependance*
- *Post-translational modifications*
- *Allosteric modifications*
- *Different kinds of interactions*
- *Compartimentalisations*
- *Quantitative aspects*
- *Role of time*

Z-conjunction: $\&_i$

Z-interaction: $\odot_i$

Z-conditional: $\rightarrow_i$

The shift from plain formulas, such as A, B, C, etc., to *labeled* formulas, such as A:**x**, B:**y**. C:**z**. etc. – where **x**, **y**, **z** are labeling strings, consisting of suitable variables, parameters or function symbols - greatly enhances the expressive power of Zsyntax and is well-grounded in contemporary logic.

The basic idea is that of generalizing the inference rules to deal with labeled formulas in which the labels are used to specify any kind of additional information concerning the entities to which the formulas refer.

Then, the purely logical mechanism is integrated with a **labeling algebra** specifying the way in which the values of the parameters should be propagated by each application of the rule.

The labels are strings of suitable variables, parameters or function symbols.

For example, EVF's can be replaced by more precise *empirically valid rules* such as:

$$\frac{A{:}[x,t_0] \qquad B[y,t_0]}{A \odot B{:}[g(x,y),t]}$$

so that, in this specific case, the corresponding instance of the (logically valid) rule of MP would take the form:

$$\frac{A \rightarrow A \odot B{:}[y,t_0] \qquad A{:}[x,t_0]}{A \odot B{:}[g(x,y),\ t]}$$

The logical content is expressed by the formulas.

The empirical content is expressed by the labels.

The general format of the labeled rules is as follows:

$$\frac{C_1 : [\mathbf{z}_1], \ldots, Cn : [\mathbf{z}_n], \quad \{A : [\mathbf{x}]\} \quad \vdots \quad B : [h(\mathbf{z}_1, \ldots, \mathbf{z}_n, \mathbf{x})]}{\overset{h}{A \rightarrow B} : [\mathbf{z}_1, \ldots, \mathbf{z}_n]}$$

The assumption in curly brackets is "discharged"

***h* is a function that depends on the the path from A to B (given C_1,,…,C_n)**

$$\frac{\overset{h}{A \rightarrow B} : [\mathbf{x}] \qquad A : [\mathbf{y}]}{B : [h(\mathbf{x}, \mathbf{y})]}$$

This intelim rules are "harmonic" (in the proof-theoretical sense): the elimination rule can be "justified" on the basis of the introduction rule.

Theorem

$TP53[a,t_0] \;\&\; MDM2[b, t_0] \;\&\; U[c, t_0] \;\&\; P[d, t_0] \vdash d(TP53)[e, t_f]$

Demonstration

1. $MDM2[a, t_0] \;\&\; TP53[b, t_0] \rightarrow MDM2 \odot TP53[l, t_1]$
2. $(MDM2 \odot TP53)\,[l, t_1] \;\&\; U[c, t_1] \rightarrow (MDM2 \odot TP53) \odot U[m, t_2]$
3. $(MDM2 \odot TP53) \odot U[m, t_2] \rightarrow MDM2[n, t_3] \;\&\; (TP53 \odot U)\,[o, t_3]$
4. $(TP53 \odot U)\,[o, t_3] \;\&\; P[p, t_3] \rightarrow (TP53 \odot U) \odot P[q, t_4]$
5. $(TP53 \odot U) \odot P[q, t_4] \rightarrow d(TP53)\,[e, t_f] \;\&\; U[u, t_f] \;\&\; P[v, t_f]$

OPEN ACCESS Freely available online

PLoS one

Zsyntax: A Formal Language for Molecular Biology with Projected Applications in Text Mining and Biological Prediction

Giovanni Boniolo[1,2], Marcello D'Agostino[3], Pier Paolo Di Fiore[1,2,4]*

1 IFOM, Istituto FIRC di Oncologia Molecolare, Milano, Italy, 2 Dipartimento di Medicina, Chirurgia ed Odontoiatria, Università di Milano, Milano, Italy, 3 Dipartimento di Scienze Umane, Università di Ferrara, Ferrara, Italy, 4 Istituto Europeo di Oncologia, Milano, Italy

Abstract

We propose a formal language that allows for transposing biological information precisely and rigorously into machine-readable information. This language, which we call Zsyntax (where Z stands for the Greek word ζωή, life), is grounded on a particular type of non-classical logic, and it can be used to write algorithms and computer programs. We present it as a first step towards a comprehensive formal language for molecular biology in which any biological process can be written and analyzed as a sort of logical "deduction". Moreover, we illustrate the potential value of this language, both in the field of text mining and in that of biological prediction.

The research programme

Future steps:

1) elaborating the correct formal framework for Zsyntax

1) constructing the right algorithm and software to implement our logic in a computer (both for text mining and for prediction)

2) showing the foundational relevance of the language

3) joining the logical language and the differential equations of molecular biology