Model of Double Strand Break of DNA in Logic-Based Hypothesis Finding

Barthelemy Dworkin[1,2], Andrei Doncescu[1,3], Jean-Charles Faye[3], and Katsumi Inoue[1]

[1] National Institute of Informatics Japan
[2] University of Toulouse, France
[3] Cancer Institut of Toulouse France

1 Introduction

Today the conception of artificial systems tries to imitate the natural systems by developing new concepts of reasoning able to handle a high level of heterogeneity and uncertainty. These complex systems have a dynamically evolution in term of structure and organization. In order to model and control these systems there is a need to observe and reconstruct their behaviour and make sense of large amounts of heterogeneous data gathered on various scales.
The science of complex systems (CSS) offers a theoretical framework for this holistic behavior by borrowing concepts from Statistical Physics, Dynamical System Theory, Theory of Computation and Machine Learning. The main characteristic of a complex system is the emergent property which arises from the interactions of low-level entities and which cannot be deduced from or explained by the properties of lower-level entities. One of the most complex systems are the living systems. If the biological systems cannot be reduced to the description of the properties of their elements it is the manner of their evolution which modify the characteristic of the constituting elements. Systems biology is expanding to cover almost all biology science, from proteins interaction to whole organ and organism-level biomedical studies by studying the interactions between the components of a biological system and how these interactions give rise to the function and behavior of the living system. Systems Biology thus attempts to understand a life process as a whole system rather than the collection of the parts considered separately. Understanding the interconnections among subsystems or elements involves a closed-loop thinking of causality. Tackling a complex system involves both synthesis and analysis, which roughly correspond to modeling and verification, respectively. Modeling is the process of generating mathematical or computational theories that satisfy specifications or goals. Hence, a candidate model is a hypothesis for a theory of the system, the modeling and verification involve inference methods, hypothesis generation/update and hypothesis verification. Similarly, in biology and any scientific field, development of a scientific theory consists of a continuous cycle of observations-explanation-prediction-experiments. Explanation of observed data is lead by hypothesis generation, and experiments lead to data that could test inferred hypotheses. Over the course

of this cycle, scientific models are frequently revised whenever discrepancies are encountered between observed and predicted results. In this paper we propose an integrated framework for reasoning on (partially observable) signaling pathway, possibly in presence of global inconsistencies

2 Double strand break of DNA

Cell's response to double strand break of DNA (DBS) has been studied for some years, but the ATM-dependant signaling pathway has only been clarified since the discovering of H2AX (Paull et al, 2000), the phosphorylated form of histone H2AX. All the protein interactions of this pathway have been reported (Pommier et al, 2005), including the signalization of the double-strand break (involving important proteins such as: H2AX, MDC1, BRCA1 and the MRN complex) but also for the checkpoints mechanisms (involving p53, the Cdc25's and Chk2). In a general way, the cell can receive information by protein interactions that will transducer signals. First, the information is discovered by sensor proteins, which will recruit some other mediator proteins whose function will be to help all the interactions between the sensors and the transducers. These transducers are proteins that will amplify the signal by biochemical methods such as phosphorylation. In the end, the signal will be given to effectors that will engage important cell process. In this pathway, the DSB is recognized by the MRN complex, which in turn will recruit ATM in its inactive dimer form, and then ATM will phosphorylate itself and dissociate to become an active monomer. This active form of ATM will phosphorylate many mediators such as ?H2AX, MDC1, BRCA1 or 53BP1. Then, the signal is transduced by important proteins such as Chk2, p53 (a very important protein, which can cause cancer if mutated) or the Cdc25's. The effectors can be different with the context: p21 and Gadd45 will induce the cycle arrest, whereas Box, Nas, Puma and Fas will induce the cell apoptosis.

3 Ampliative Reasoning in Biological Systems

Understanding genetic and metabolic networks is of the utmost importance. These networks control essential cellular processes and the production of important metabolites in microorganisms. Since Metabolic Control Analysis teaches us that control over metabolic flux is shared between the different enzymes in a pathway, it follows that it is also easier to change metabolic fluxes (and product yields) by modulating the regulator that is in charge, rather than by separately adjusting the activity of each enzyme in the pathway. Modeling such networks from model organisms will drive applications to other less characterized organisms, which have a high biotechnological potential. Furthermore, such modeling is a precondition for the growth of synthetic biology, i.e. building up "newly created biological systems" which will generate unknown products from simple organisms and should have a strong economical impact in the future. The logical approach provides an intuitive method to provide explanations based on an

expressive relational language. For example, logic can represent biological networks such as gene regulatory, signaling transduction, and metabolic pathways. Unlike other approaches, this method allows a background theory, observations and hypotheses within a common declarative language, and provides the basis for the three main forms of inference, i.e., deduction (prediction), abduction (explanation) and induction (generalization). Although deduction produces logically correct (sound) consequences of the theory, both abduction and induction provide expansion of the logical theory, that is, they augment the original theory by adding new hypotheses. Note that deduction has traditionally been used for proving theorems of a given axiom set, but here we need to find new consequences, i.e., consequence finding, which is more general than theorem proving. Interestingly, the hypothesis-finding problem (abduction and induction) can be translated into consequence-finding problems, so that we can realize all three modes of inference using a deductive, consequence-finding procedure. We mention the Inductive Inference as a type of reasoning that justifies some modification from one state of absolute believe to another by adding new information to the initial assumes that is consistent with it but does not entail it. Therefore, Inductive Inference is ampliative in the Pierce meaning and non monotonic.

3.1 Hypothesis Finding from First-order Full Clausal Theories

The importance of hypothesis generation has been more and more recognized recently for many innovative applications. SOL and SOLAR [Siegel, Inoue] is the state-of-the-art consequence-finding system in this field, whose performance is comparable with high-speed theorem provers when applied to theorem proving, and is the only system which is sound and complete for consequence finding in full clausal theories. Abduction using SOLAR] and CF-induction [Ino04] are also the unique abductive and inductive reasoning systems, respectively, which are sound and complete for first-order full clausal theories. None of other abductive or inductive systems can be comparable with them.

We will first present the case study and some behavior rules. Then, we will present the limits of the classical logic and why we need the nonmonotonic logic. We will explain the formalization of the behavior rules with default logic. We use only normal defaults and Horn clauses in order to simplify the program, though we could extend this work to other case studies, with more complicated rules. Next we explain the choice between the extensions thanks to preferences with simple probabilistic techniques.

4 Logical Model of Double Strand Break

Transfer reactions such as phosphorylation (and autophosphorylation), ubiquitination and methylation are specific cases of enzymatic stimulation. The mechanism is simple: an enzyme take a substrate and make a covalent bound with a marker. We can simplify this transfer reaction like this:

$$Reaction(enzyme, substrate) \Rightarrow product(marked - substrate) \quad (1)$$

In details, phosphorylation is a transfer of an inorganic phosphate (most of the time brought by an energy-bearing molecule such as ATP) and autophosphorylation is a special case where a protein can phosphorylate itself.

$$Phosphorylation(enzyme, substrate) \Rightarrow product(P - substrate) \quad (2)$$

Ubiquitination is a transfer of a small peptide called ubiquitin and methylation is a transfer of a methyl component (CH3).

1. $Ubiquitination(enzyme, substrate) \Rightarrow product(Ub - substrate)$
2. $Methylation(enzyme, substrate) \Rightarrow product(CH3 - substrate)$

Unlike the other reactions, the transcription activation is not a transfer reaction, but a very important one though. It means that a protein will bind on a specific location of DNA and will induce the transcription of the target gene in RNA (along with a complex of others important proteins). After translation of the mRNA, a raw peptide will be produced and will be eventually modified by post-translational reactions such as phosphorylation, methylation or even glycosylation, which will give the protein related to this gene. We decided to model the transcription activation this way:

$$Transcriptionactivation(transcriptionfactor, promoter) \Rightarrow product(translatedprotein) \quad (3)$$

5 Results

This logical model contains two significant points for biological applications. On the first hand, nine different predicates are used to describe the biological interactions: enzymatic stimulation (general or not precise), phosphorylation, *autophosphorylation*, *ubiquitination*, *binding*, *transcriptionactivation*, *dissociation* and *product*. We precise that *phosphorylation*, *autophosphorylation*, *ubiquitination*, *transcriptionactivation* are just specific cases of enzymatic stimulation. The binding and dissociation are opposite reactions: the binding can add two proteins to get a product, whereas the dissociation divides this product into two new products - and thus needs two clauses, one for each new product. The 'product' predicate describes the production of a protein following a reaction. It can be surprising but we did not choose to include a predicate for inhibition reactions. The fact is that in biology, 'inhibition' does not always describe the same mechanism: it can be an inhibition of a protein induced by binding with another protein, or another way to say 'if A exists, then B won't exist' (where A and B could be proteins or pathways). So, in order to still include the inhibition in our model without a specific predicate, we decided to find another way: so instead of

the clause $product(A) \Rightarrow inhibition(A, B)$ we use $product(A) \Rightarrow (product(B))$. This method has revealed an good potential, and can be checked. Here is an example: normally in the cell, the protein $Cdc25A$ exists and prevents the cell cycle arrest. But if this protein is phosphorylated (by $Chk1$ for instance), it will be recognized by degradation effectors and will be dismantled, and without $Cdc25A$ the cell cycle stops. In the data, we modeled all the information this way:

1. $stimulation(cdc25a, cell) \Rightarrow (product(cellcyclearrest))$
2. $product(pchk1) \Rightarrow phosphorylation(pchk1, cdc25a)$
3. $phosphorylation(pchk1, cdc25a) \Rightarrow product(pcdc25a)$
4. $product(pcdc25a) \Rightarrow stimulation(pcdc25a, cdc25adegradationeffectors)$
5. $stimulation(pcdc25a, cdc25adegradationeffectors) \Rightarrow product(cdc25adegradation)$
6. $product(cdc25a_degradation) \Rightarrow (stimulation(cdc25a, cell))$

This is the top clause and the production filled used to test the efficiency of our inhibition:

1. cnf(tp1, topclause, [-stimulation(cdc25a,X),-product(cellcyclearrest),ans(X)]).

2. pf([-stimulation(cdc25a,),-product(cellcyclearrest),ans( )]).

In biological words, this means we want to ask SOLAR to find all the substrates of Cdc25A (by a simple stimulation) that will induce the cell cycle arrest. And the only answer is:

$$conseq([+ans(0), -stimulation(cdc25a, 0), -product(cellcyclearrest)]). \quad (4)$$

No other consequence found, but it is normal, for Cdc25A prevents the cell cycle arrest. Then, we ask SOLAR to find all the substrates of the phosphorylated Cdc25A on the same method:

1. cnf(tp1, topclause, [-stimulation(pcdc25a,X),-product(cellcyclearrest),ans(X)]).
2. pf([-stimulation(pcdc25a,),-product(cellcyclearrest),ans( )]).

And here are the answers:

1. conseq([+ans(cdc25adegradationeffectors), -product(cellcyclearrest)]).
2. conseq([+ans(0), -product(cellcyclearrest), -stimulation(pcdc25a, 0)]).

It tells us that the Cdc25A degradation effectors will induce the cell cycle arrest, which is true because if $Cdc25A$ is brought into degradation (after is phosphorylation), the cell cycle is stopped. In conclusion, our inhibition method by not creating a specific predicate is efficient.

We made an experiment to see if our model could fix relations between proteins if we excluded a single protein. In a simple example of 4 proteins connected to each other by a series of simple implications $A \Rightarrow B \Rightarrow C \Rightarrow D$, we wanted

to prove that our model could make a relation from B to D if C was excluded. Here is the case of interactions with $RNF8$:

In biological words, $RNF8$ binds with the phosphorylated form of $MDC1$, thus allowing $UBC13$ to bind on $RNF8$. And finally, the complex $RNF8/UBC13$ can do an ubiquitination on the $H2A$ histone, thus creating the ubiquitinated form $Ub-H2A$. We decided to delete the $RNF8/UCB13$ complex to see if $SOLAR$ could find a way to make a relationship between $RNF8-bound$ and $Ub-H2A$.

In our model, we structured the data biological data into SOLAR:

1. cnf(rnf-02, axiom, [-binding(p-mdc1,rnf8),product(rnf8-bound)]).
2. cnf(rnf-03, axiom, [-product(rnf8-bound),binding(rnf8-bound,ubc13)]).
3. cnf(rnf-04, axiom, [-binding(rnf8-bound,ubc13),product(rnf8-ubc13)]).
4. cnf(rnf-05, axiom, [-product(rnf8-ubc13),ubiquitination(rnf8-ubc13,h2a)]).
5. cnf(rnf-06, axiom, [-ubiquitination(rnf8-ubc13,h2a),product(ubc-h2a)]).

Then we deleted the predicate product(rnf8$_u$bc13)$whichwaspresentintwoclauses$.

1. cnf(rnf-02, axiom, [-binding(p-mdc1,rnf8),product(rnf8-bound)]).
2. cnf(rnf-03, axiom, [-product(rnf8-bound),binding(rnf8-bound,ubc13)]).
3. cnf(rnf-06, axiom, [-ubiquitination(rnf8-ubc13,h2a),product(ub-h2a)]).

The idea was to prove that $RNF8_bound \Rightarrow Ub-H2A$ was true. We wanted to search all the products that would lead to the production of $Ub-H2A$ in order to see if $RNF8-bound$ was among these products. We used this top clause and this production field:

1. $cnf(tp-1, top_clause, [-product(X), -product(ub-h2a), ans(X)])$.
2. $pf([-product(), -product(ub-h2a), ans()])$.

This experiment was tested on a computer under Windows 7, with an *intel* Core i5 $2.66GHz$ processor and 4096 of RAM. It lasted 163 seconds. As an answer, 527 new consequences have been found (with no time limit and no other parameters). Among these consequences, we found three of them that had the predicate product($rnf8-bound$):

1. conseq([+ans(rnf8-bound), -product(ub-h2a), -product(p-mdc1)]).
2. conseq([+ans(rnf8-bound), -product(ub-h2a), -product(h2ax-mdc1)]).
3. conseq([+ans(rnf8-bound), -product(ub-h2a), -product(gamma-h2ax)]).

These results show that as we expected, $RNF8-bound \Rightarrow Ub-H2A$ is true, and the production of the ubiquitinated histone is the consequence of the production of $H2AX$, of $MDC1$ bound to $H2AX$ and of the phosphorylated form of $MDC1(P*-MDC1)$.

6 Conlusion

We propose a complete logical model, respecting the protein interactions and biological veracity. Its first use is to act like a database: for instance, a biologist working on this pathway can ask all the proteins phosphorylated by Chk2, or all the proteins necessary for the ubiquitination of HA2 histone by RNF8 and UBC13. Depending of the SOLAR top clause, the question may be simple or

complicated in biological terms. The other hypothetical use of our model is to give information of a new protein that would interact with proteins from this pathway. For instance, if a new protein is discovered and known to interact with a specific protein of the cell response to DSB pathway, our model could find consequences of this interaction.

References

1. K. Inoue: Induction as consequence finding. Machine Learning, 55:109–135, 2004.
2. Kitano H.: Systems Biology Toward System-level Understanding of Biological Systems Kitano. In Science Vol. 295. no. 5560, pp. 1662-1664 (2002).
3. R. J. Mooney: Integrating abduction and induction in machine learning. In Working Notes of the IJCAI97 Workshop on Abduction and Induction in AI, 37–42 (1997). luwer Academic Press
4. Geurts P.: Pattern extraction for time-series classification. Proceedings of PKDD 2001, 5th European Conference on Principles of Data Mining and Knowledge Discovery, LNAI 2168, 115-127 (2001).
5. Keogh, E. and Lin, J. and Fu, A.: HOT SAX: efficiently finding the most unusual time series subsequence. Fifth IEEE International Conference on Data Mining, 8 pp.- (2005).
6. Schwarz G.: Estimating the dimension of a model. Annals of Statistics, Vol.6, No.2, pp.461-464 (1978).
7. Cheeseman P. and Stutz J.: Bayesian classification (AutoClass): Theory and results. Advances in Knowledge Discovery and Data Mining, pp.153-180, The MIT Press (1995).
8. Beal M. J.: Variational Algorithms for Approximate Bayesian Inference. PhD. Thesis, Gatsby Computational Neuroscience Unit, University College London (2003).
9. Gauvain J.-L. and Lee C.-H.: Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Transactions on Speech and Audio Processing, Vol.2, Issue 2, pp.291-298 (1994).
10. Ji S., Krishnapuram B., and Carin L.: Variational Bayes for continuous hidden Markov models and its application to active learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.28, Issue 4, pp.522-532 (2006).
11. Kanehisa, M. and Goto, S.; KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27-30 (2000).
12. Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y.; KEGG for linking genomes to life and the environment. Nucleic Acids Res. 36, D480-D484 (2008).
13. Nabeshima H., Iwanuma K., and Inoue K.: SOLAR: A Consequence Finding System for Advanced Reasoning. Proceedings of the 11th International Conference TABLEAUX 2003, Lecture Notes in Artificial Intelligence, Vol. 2796, pp. 257-263, Springer (2003).