Investigating cellulose degradation: placing Qualitative Reasoning in the process

Kamal Kansou
INRA, UR 1268
Biopolymères Interactions & Assemblages
BP 71267
44316 Nantes Cedex 3, France
kamal.kansou@nantes.inra.fr

Bert Bredeweg
Informatics Institute, University of Amsterdam
Science Park 904
1098 XH Amsterdam, The Netherlands
B.Bredeweg@uva.nl

ABSTRACT

Scientific research produces a vast volume of information and knowledge about natural phenomena, typically published in papers. This is particularly striking for the enzymatic hydrolysis of lignocellulose, a critical bioprocess for the production of second-generation biofuel. Our objective is to build Qualitative Reasoning (QR) models that capture the knowledge reported in scientific papers and implement putative explanations for concrete observations. QR is an Artificial Intelligence modelling technique that captures knowledge as causal relations to simulate the system behaviour over time from its structure. The rationale for using a qualitative over a quantitative technique is mainly the incomplete understanding of a system, in this case the cellulose degradation mechanism. When developing a QR model of this kind, we first create a base-model, which is then extended to included more features, and explain additional observations. The model presented in this paper captures the interpretations described in three different scientific papers related to the target system and its behaviour. The base-model implements an interpretation based on the accumulation of inactive enzymes. The extension contains model fragments that capture knowledge about the substrate conditions over the process. Both the capacity to represent results of each paper and the target behaviour are examined and discussed.

Categories and Subject Descriptors

I.2.1 [**Artificial Intelligence**]: Applications and Expert Systems—*Medicine and science*

General Terms

Keywords

Knowledge modelling, Biological system, Enzymatic reaction, Qualitative reasoning

1. INTRODUCTION

The conversion of plant waste into sugar and then into energy using enzymes is a strategic bioprocess for the production of biofuel. The potential environmental impact boosts the domain research that produces an abundant amount of scientific literature. Because the process is both complex and incompletely understood, a large number of factors is investigated under a wide range of experimental conditions. This affects the expression and the assessment of theories about the underlying processes. Fundamental tasks are becoming problematic, such as retrieving information, confront or associate interpretations from distinct papers to promote cross-fertilization. Designing intelligent programs regarding this issue is an important challenge for KR and AI, and the construction of explanatory models will be a part of it.

Well suited representations for building explanatory models are available in the Qualitative Reasoning (QR) framework. QR proceeds from descriptions of interacting components or processes in a symbolic, human-like manner [10]. Some QR approaches provide effective means for capturing knowledge as causal relationships and producing dynamic simulations to envision how these causal relationships determine the system behaviour. QR uses qualitative abstraction to capture relevant aspects of a system without the need of precise numerical data, which makes this approach suitable for modelling systems with partial and imprecise information.

In this paper we describe the construction of QR model representing an integrated system originating from interpretations taken from three scientific papers. We also reflect on methodological issues relevant to creating such a model. The QR model is innovative in the sense that it generates a plausible explanation for a target behaviour that extends the knowledge chunks elicited from the three source papers. Our paper demonstrates the relevance of QR to capture and assemble scientific knowledge from different documents and build a new scientific interpretation.

2. QUALITATIVE REASONING

Qualitative Reasoning (QR) is an area of Artificial Intelligence that strives for inferring *behaviour from physical system structure* in a symbolic, human-like manner. The firm causal and mathematical foundation of the QR approaches [19] guarantees the soundness of the automatic reasoning generating the simulation results. QR proceeds from descriptions of interacting entities (representing the physical system structure), quantities and processes, and is informed by research in cognitive science about how humans reason [3]. QR modelling combines qualitative abstraction to lower precision of the system quantities and the ability to capture conceptual knowledge such that it represents explanations of phenomena.

In this study we use Garp3 [4], it is a workbench for constructing and simulating qualitative models. It also provides tools for inspecting the simulation results. The ontology and the formalism provided by Garp3 derive from typical approaches to qualitative reasoning [2]. A Garp3 model involves several ingredient types. Entities are the structural elements endogenous to the system. Quantities are the properties of entities characterized by: <*Magnitude, Derivative*>. The domain of magnitudes for a quantity is called the Quantity Space. It is a finite and discrete set of symbols. A quantity space is an abstraction or a mapping of a continuous numerical scale consisting of a succession of alternating points (also called landmarks) and intervals. All the derivatives have by definition the same quantity space, namely: {min, zero, plus}. The key cause-effect relationships are: Direct Influence (I+; I-) and Indirect Influence or qualitative proportionality (P+; P-). The former represents the cause of the changes, whereas the latter represents the propagation of these changes [9]. Garp3 reasoning engine integrates value correspondences between quantity values and reasoning on inequalities and algebraic relations that act as constrains.

2.1 Related work

It is one of the traditional use of QR languages to model domain theories, initially in physics [8, 5]. Many QR models have been developed in ecology [18, 15], others can be found in other domains such as social science [13, 12]. Most of those studies focus on capturing existing domain theories as an explicit knowledge model to be used to convey explanation about a phenomena to other scientists, students or stakeholders. Forbus' Qualitative Process Theory [9] and Garp3 are often used for that purpose. The process-ontology that these inference engines are based on matches the way specialists reason about systems whose structure is partially unknown.

Automatic identification of models is another dynamic area of research for which QR approaches have been developed. Typical applications concern the knowledge discovery regarding metabolic pathways and genetic networks [7, 16]. Work in this domain usually deploys a version of QR that uses a qualitative model in the form of Qualitative Differential Equations (QDEs) to envision the behaviour of a system [17].

Our work, as presented in this paper, takes inspiration from both types of application for identifying the plausible integrated causal explanation for the ideas described in the three scientific papers.

2.2 Enzymatic hydrolysis of cellulose

Cellulose is a major component of the plant cell wall and the largest and accessible renewable source of carbon on Earth. It is an insoluble linear macromolecule, composed as a chain of glucose units. Cellulose is degraded by cellulases, i.e. enzymes degrading cellulose, to produce glucose or small chains of glucose, such as cellobiose (2 glucose units). The most important cellulase in this regard is cellobiohydrolase (CBH), which, once complexed with the cellulose, digests the cellulose strand in a processive manner releasing cellobiose units at each catalytic step until it desorbs to get back in solution. Experimentations about this reaction involve a cellulosic substrate, that can be of many kinds, and an enzymatic system, ranging from a single cellulase, usually CBH, to an enzymatic cocktail. The partially unknown structure of cellulose-enzyme(s) systems explain why building mechanistic model predicting the degradation kinetic is challenging. In particular both enzyme and substrate-related factors can be held responsible for the conversion rate limitation. Most models proposed in the literature are kinetic models, based on semi-mechanistic considerations [1].

We chose to study the enzymatic degradation of a cellulose with a commercial of enzymatic cocktail for a long time (∼150 hours). This is the classical experimental condition reported in many papers. Yet, if one asks about prominent aspects of the resulting progress-curves, for instance reaction rate slowing-down or fraction of recalcitrant substrate, domain specialists encounter difficulties in providing definite explanations. Hence, our modelling challenge is the following: *can we provide a plausible explanation for this target behaviour, with scientific domain literature as our primary knowledge source?*

3. DESIGNING SELF-EXPLANATORY MODELS FROM PAPERS

Using a QR language, it is possible to capture indistinctly the causal relations about a given process from a group of papers. The computational complexity of a QR model built with such an approach will quickly make the simulation intractable and inappropriate to convey a meaningful explanation to domain specialists. Instead, we adopt a progressive approach. Driven by a target behaviour, we strive for being very selective about the knowledge sources and about the processes to be modelled. Thus a process not directly involved in the explanation is either left aside or represented in an abstract way (if deemed necessary to get a complete system).

Taken this approach work, the first task is to define a target of interest for the experts, and of reasonable size for the QR model. The target is actually a qualitative system that exhibits one or more target behaviour(s). The target behaviour is a qualitative abstraction of observed behaviours exhibited by a target system. It defines the qualitative features of the observed behaviours that have to be explained. In doing so, the modeller determines the quantities (i.e. the variables in a QR model) and the quantity spaces relevant for simulating the observed behaviours [14]. In the ideal case, target behaviour is a straightforward mapping of existing data. However, in natural sciences the dataset at hand might not be informative enough, for instance due to costly experimentations. Qualitative abstraction reduces the distinction between the experimental results from different papers, as a result a large spectrum of published materials can be used to enrich the target behaviour.

Our modelling methodology is depicted in Fig. 1. The QR model is built incrementally by capturing the interpretation from at least one scientific paper (usually in the discussion section). Selected papers display observations or simulation results describing processes related to the target and may give useful interpretations. Each knowledge chunk is captured as a model fragment of the QR model. This exploits the compositional modelling feature of Garp3 [4]. For each version of the QR model that conveys a candidate explana-

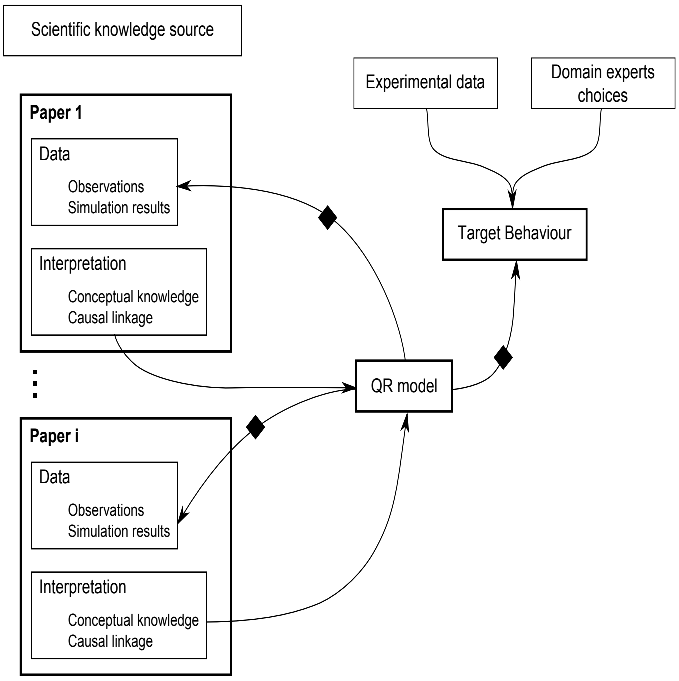

Figure 1: Diagram of the approach adopted for building a QR model from domain literature. Link with black diamond represents test of simulation *vs* observations

tion two criteria are assessed (Fig. 1):

Encompassment The QR model is a consistent representation of the interpretations given in the source papers. The model generates behaviours that match the observed data, numerical simulations or qualitative observations supplied in these papers.

Sufficiency The QR model implements a plausible explanation for the target behaviour. The model generates a behaviour from which a plausible explanation for the target behaviour can be derived.

Definition of the target behaviour and selection of source papers are critical stages that determine the content and then the properties of the explanatory model. Both are carried out jointly with the domain experts.

4. MODELS OF CELLULOSE ENZYMATIC HYDROLYSIS

4.1 Defining the target

Hydrolysis of cellulose is characterised by progress-curves of the degradation of cellulose into smaller molecules. The curve exhibits a general saturation-shape and reflects that reflects the enzyme action. The hydrolysis rate gradually slows-down with time. This problem is actively investigated as it limits the conversion efficiency. The main objective is to propose an explanatory model for the rate slowing-down from existing knowledge.
The Target Behaviour is a composite object built from concrete experiments and supplemented with observations from the literature. Experiments are 3 hydrolyses of cellulose Avicel (processed crystalline cellulose) from 168h to 216h with 2 initial concentrations of a commercial enzymatic cocktail 50 mg/g and 10 mg/g. The experiments consist in adding fresh enzymes in the course of the process to see if this boosts (we call onward re-start) the hydrolysis process. Three modalities of enzyme concentration are tested:

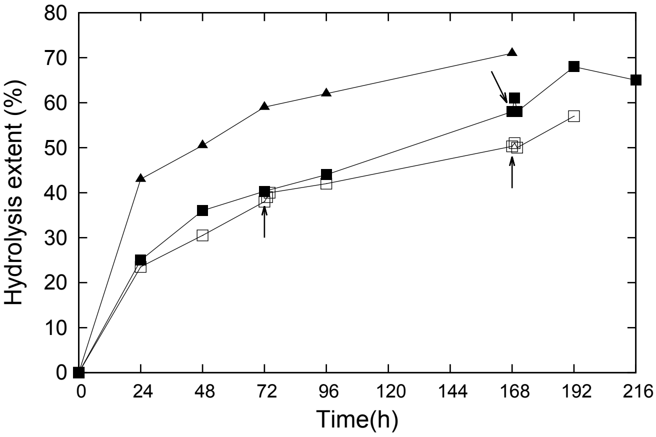

Figure 2: Hydrolysis curves of cellulose Avicel for 3 concentrations of enzymes and addition of enzymes at different times, indicated by arrows. ▲ 50 mg/g, ■ 10 mg/g and addition t=168h,□ 10mg/g and addition t=72h and t=168h.

- Initial concentration of 50 mg/g, no addition of fresh enzymes
- Initial concentration of 10 mg/g, addition of 50 mg/g fresh enzymes at t=168h
- Initial concentration of 10 mg/g, addition of 10 mg/g fresh enzymes at t=72h, and 40 mg/g at t=168h

The progress-curves, Fig. 2, depict two phases, a rapid phase during the first hours of the reaction, and from 24 hours to 168 hours a slower seemingly linear phase, with comparable hydrolysis rates whatever the enzyme concentration in the solution. In Fig. 2 no clear re-start is observed. To enrich this description domain experts selected 3 types of experiments described in the literature. Given the experiments, Fig. 2, and additional observations, the expected behaviour has the following features:

- Addition enzyme. No restart
  - Surface clean-up & Addition enzyme. Restart
- Addition substrate. Restart
- Hydrolysis rate slowing-down. No or weak dependency on initial enzyme concentration

The two first experiments are obviously related as the surface clean-up is informative only when no restart was observed. The paper focuses on the Addition enzyme property.

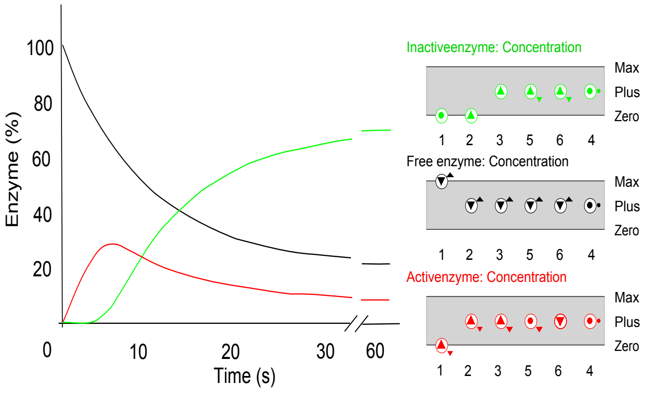

Figure 3: Left graph is reproduction of simulation results from [6], graphs on the right illustrate the corresponding behaviour pathway produced by the QR base model.

4.2 Establishing the base model

Cruys-Bagger et al., [6] paper is published in a journal of biochemistry. It is a kinetic study for hydrolysis of processed amorphous cellulose by a cellobiohydrolase (CBH) for 1 minute. The goal of the paper is to investigate the slowing-down of the hydrolysis rate for the first minute of the reaction through the identification of the rate-limiting factors. The authors present a mechanistic kinetic model of the enzymatic reaction to support their interpretation of the results. The study focuses on fundamental processes that occur in the target system as well. The authors identified a strong hydrolysis rate slowing-down almost at the onset of the reaction from the following model:

$$\begin{array}{llll}
E + C_m \xrightarrow{k_{on}} & EC_m \xrightarrow{k_{cat}} & EC_{m-1} \cdots \xrightarrow{k_{cat}} & EC_{m-n} \\
k_{off}\downarrow & k_{off}\downarrow & & k_{off}\downarrow \\
E + C_m & E + C_{m-1} & & E + C_{m-n}
\end{array}$$

E represents the CBH that adsorbs on the cellulose surface at a reaction rate constant k_{on}, hydrolyses a cellulose strand, represented by C, in a processive manner at k_{cat}, to release cellobiose. A cellulose strand is composed of m cellobiose units, it is assumed that on average n units of cellobiose are released by a CBH before it gets stalled by some obstacles at the cellulose surface as EC_{m-n} complex. In this form the enzyme needs to dissociate at k_{off}. Using the model the author produces the simulation curves (Fig. 3), where 70% of enzyme are stalled and then inactive after 1 min, limiting the quantity of active enzyme. The rate limitation is related to the accumulation of inactive enzyme, due to morphological obstacles (low n) and slow dissociation velocity (low k_{off}).

Fig. 4 presents an overview the corresponding QR model structure, produced automatically by Garp3 for the initial state of the simulation. It depicts the relations between the quantities of the model. Entities of the model are boxes linked via semantic relations, not represented in Fig. 4 for clarity sake. The QR model has three main entities: Cellobiohydrolase (Enzyme), Cellulose (Substrate), Cellobiose (Product). The enzymes can be in 3 states: (i) Free in solution, (ii) Active during the production stage, (iii) Inactive. Quantities assigned to the entities are in the boxes. In the base model it includes *Concentration* and rates. Contrary to the kinetic model, the base model does not include the three first-order constants but the corresponding rates, *Rate on, Rate cat, Rate off*. *Concentration* in Free enzyme stands for E, *Concentration* in Inactive enzyme stands for EC_{m-n} complex, while *Concentration*in Active enzyme stands for EC_{m-i} *with* $i \in \{0, \ldots n-1\}$. The QR model, like the kinetic model, captures exclusively the relations between the concentrations of enzyme in the different states. Neither the substrate nor the product play an active role in the base model. The position of the active enzyme along the cellulose strand is not captured in the base model. While determinant in the kinetic model to obtain realistic simulations, it is not needed to convey interpretation of the results using QR. Quantity spaces of the Quantities are given below the Quantity label current value in red and derivative sign as symbols {▲, ●, ▼} for decrease, steady or increase.

Running the simulation produces a state-graph of 27 qualitative states with a characteristic fan shape (Fig. 4b). State 4 is the dynamic equilibrium state, with all quantities of the system steady but the concentration of cellobiose that increases at a constant rate. The simulation depicts two kinds of behavioural pathways of interest. One maps the kinetic model simulation (Fig. 3 and Fig. 4b in red). Fig. 3 displays a perfect match between the two simulations. Hence, the base model shows *encompassment* for paper [6] as it can convey the interpretation of the results in the form of causal graph (Fig. 4a). It appears that the kinetic simulation is described by a pathway of 6 qualitative states. For each state, a causal graph similar to Fig. 4a is produced. Thus, state 4 (Fig. 3), displays the accumulation of Inactive enzyme affecting the hydrolysis.

The second behaviour is cyclic and goes around state 4, from state 3 via states 9, 17 and finally 26 until it meets state 3 again (Fig. 4b). State 4 can be reached from many states of this pathway. The behaviour depicts successive oscillations of the 3 concentrations of enzymes, starting from Active enzyme, as captured by the chunk [3→5→6] (Fig. 3). While this alternative behaviour does not match the kinetic model simulation, it seems like a valid physical description. Indeed, if one considers enzyme as discrete agents and not as a continuous and innumerable quantity, then the equilibrium state (state 4) will never be exactly reached but instead the number of enzymes in the 3 states will oscillate around it.

4.3 Confrontation with the target behaviour

Cruys-Bagger et al.,'s system clearly does not exhibit the target behaviour. Consider the simulation results (Fig. 3). There is less free enzymes in the solution as a significant part of it, is inactive (state 4). If one adds fresh enzyme at that moment, then following the system (Fig. 4) there will be more Free and then Active enzyme, so the hydrolysis will restart[1] and the following proposition applies:

- Substrate available ∧ Enzymes not functioning → Addition enzyme. Restart

This however contradicts the target behaviour. Therefore if enzyme is added after several hours of hydrolysis the expla-

[1] Cruys-Bagger et al., actually did the test by adding the same amount of enzyme after t=60 sec and observed two bursts of hydrolysis of comparable magnitude

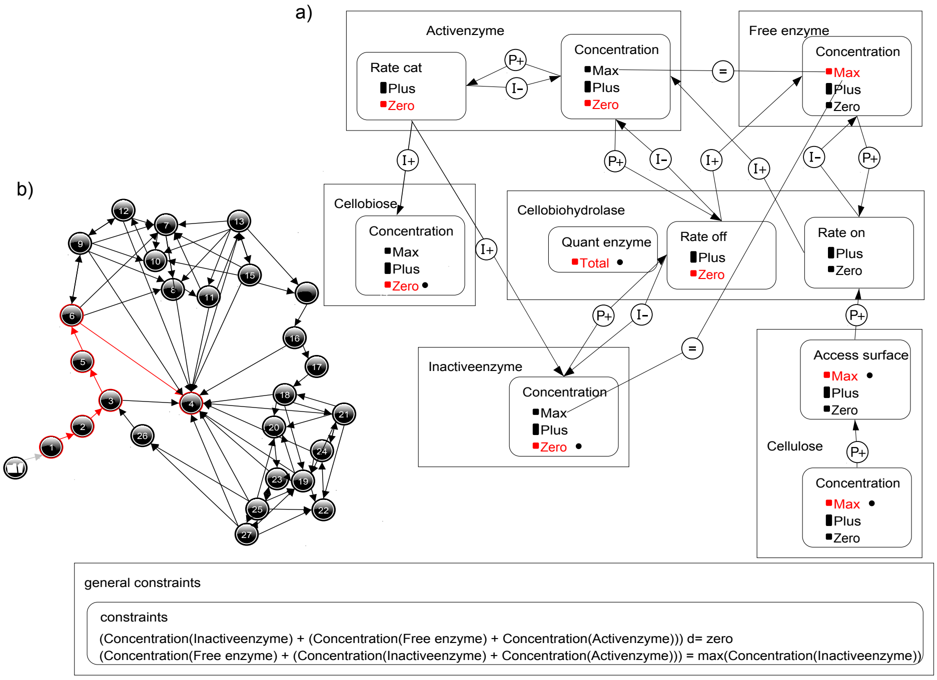

Figure 4: QR base model simulation results: (a) causal graph underlying state 1, (b) the state-graph.

nation conveyed by the base model gets insufficient. There is a need to understand why.

4.4 Integrating the substrate condition

Fox et al., [11] paper is published in a journal of biochemistry. It is a kinetic study for hydrolysis of BMCC (bacterial microcrystalline cellulose) for 100h, much longer than the paper [6]. Their findings allow them to propose that the rate of complexation of CBH with the cellulose limits the rate of CBH-catalyzed hydrolysis. In the QR model *Rate on* represents both the rate of complexation and the rate of adsorption. A limiting *Rate on* contradicts apparently Cruys-Bagger et al.,'s interpretation based on a low *Rate off.* A third input is needed to relate these interpretations. Yang et al., [20] paper is published in a journal of biotechnology and bioengineering. It is a study of cellulose Avicel hydrolysed by a complete cellulase system during 15h with different restart experiments. The conditions described in this paper are comparable to the ones used to produced Fig. 2. Accordingly the authors observed that the addition of fresh Enzymes causes weak restart unless a cellulose surface clean-up were performed beforehand. The authors suggest that the surface of the cellulose is actually more accessible later in the reaction but it is enzyme attached to cellulose surface that affects the hydrolysis. As more enzyme adsorbs on the surface a steric hindrance might appear.
Building an executable version of these interpretations requires the modelling of the substrate surface condition. We consider that an accessible surface of cellulose can be either covered or available for the enzyme to adsorb. A new model fragment is added to implement a connection between *Concentration* of Inactive enzyme and a *Covered access surface*:

- *Concentration* [in Inactive enzyme] $\xrightarrow{P+}$ *Covered access surface* [in Cellulose]
- *Access surface* [in Cellulose] $\geq$ *Covered access surface* [in Cellulose]
- *Access surface* [in Cellulose] - *Covered access surface* [in Cellulose] = *Available access surface* [in Cellulose]
- *Covered access surface* [in Cellulose] $\xrightarrow{P-}$ *Available access surface* [in Cellulose]
- *Access surface* [in Cellulose] $\xrightarrow{P+}$ *Available access surface* [in Cellulose]

The available accessible surface can limit the adsorption of Free enzyme. This is implemented in a second model fragment as follows (P* is the proportionality relation that corresponds to the product):

- *Available access surface* [in Cellulose] × *Concentration* [in Free enzyme] = *Rate on* [in Cellobiohydrolase]
- *Available access surface* [in Cellulose] $\xrightarrow{P*}$ *Rate on* [in Cellobiohydrolase]
- *Concentration* [in Free enzyme] $\xrightarrow{P*}$ *Rate on* [in Cellobiohydrolase]

Capturing these new relations generates a state-graph of 41 states, with one dynamic equilibrium state, again state 4. The behaviour pathway of Fig. 3 is also produced, so the model maintains *encompassment* for [6]. The extended version now relates *Rate on* dynamic to the *Concentration* of Inactive enzyme via the *Available access surface.* An instance of the cyclic behaviour is particularly illustrative in

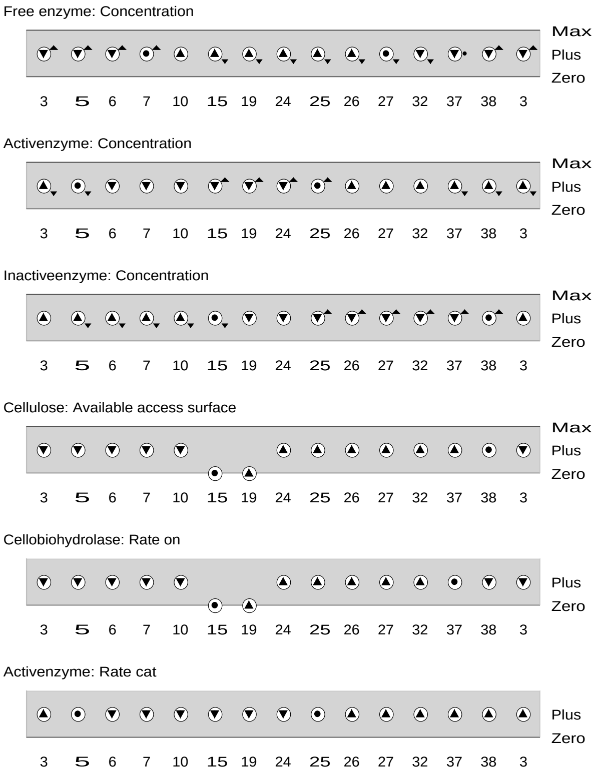

Figure 5: Value-history graph for one behaviour pathway produced with the extended QR model.

this respect (Fig. 5). The accumulation of Inactive enzyme can increase, up to cover the accessible surface and stops the recruitment of Active enzyme (*Rate on*=zero) in state 15 (Fig. 5). State 15 will last until some Inactive enzyme is released via *Rate off*, which is a slow process as discussed before. In such conditions the rate of complexation via *Rate on* might appear limiting and adding fresh enzyme without removing the Inactive enzyme from the surface will not cause a restart. This matches [11, 20] interpretations and the target behaviour for the addition of enzyme experiments.

5. CONCLUSIONS

This paper describes the construction of self-explanatory QR models as instruments to integrate ideas presented in different scientific papers. The notions of *encompassment* and *sufficiency* are postulated as criteria to evaluate the appropriateness of a particular model for a given set of papers.

Acknowledgments

The authors would like to thank the domain experts involved in this project, Jean Tayeb, Caroline Rémond, Gabriel Paës and Estelle Bonnin for their assistance in this work.

6. REFERENCES

[1] P. Bansal, M. Hall, M. J. Realff, J. H. Lee, and A. S. Bommarius. Modeling cellulase kinetics on lignocellulosic substrates. *Biotechnology Advances*, 27(6):833–848, 2009.

[2] B. Bobrow. Qualitative reasoning about physical systems: An introduction. *Artificial Intelligence*, 24(1-3):1–5, 1984.

[3] B. Bredeweg and K. Forbus. Qualitative modeling in education. *AI Magazine*, 24(4):35–46, 2003.

[4] B. Bredeweg, F. Linnebank, A. Bouwer, and J. Liem. Garp3 - workbench for qualitative modelling and simulation. *Ecological Informatics*, 4(5-6):263–281, 2009.

[5] J. Collins and K. Forbus. Building qualitative models of thermodynamic processes. In *Third International Workshop on Qualitative Reasoning Proceedings*, pages –, 1989.

[6] N. Cruys-Bagger, J. Elmerdahl, E. Praestgaard, H. Tatsumi, N. Spodsberg, K. Borch, and P. Westh. Pre-steady-state kinetics for hydrolysis of insoluble cellulose by cellobiohydrolase cel7a. *Journal of Biological Chemistry*, 287(22):18451–18458, 2012.

[7] H. de Jong, J. Geiselmann, B. Batt, H. C., and P. M. Qualitative simulation of the initiation of sporulation in bacillus subtilis. *Bulletin of Mathematical Biology*, 66:301–340, 2004.

[8] B. de Kleer and J. Brown. A qualitative physics based on confluences. *Artificial Intelligence*, 24(1-3):7–83, 1984.

[9] K. Forbus. Qualitative process theory. *Artificial Intelligence*, 24(1-3):85–168, 1984.

[10] K. Forbus. *Hanbook of Knowledge Representation*, chapter Qualitative Modelling. Elsevier, 2008.

[11] J. M. Fox, S. E. Levine, D. S. Clark, and H. W. Blanch. Initial- and processive-cut products reveal cellobiohydrolase rate limitations and the role of companion enzymes. *Biochemistry*, 51(1):442–452, 2012.

[12] N. Fridman and G. A. Kaminka. Using qualitative reasoning for social simulation of crowds. *ACM Trans. Intell. Syst. Technol.*, 4(3):1–21, 2013.

[13] J. Kamps and G. Péli. Qualitative reasoning beyond the physics domain: The density dependence theory of organizational ecology. In *QR95 Proceedings*.

[14] K. Kansou and B. Bredeweg. Hypothesis assessment with qualitative reasoning: Modelling the fontestorbes fountain. *Ecological Informatics*, 19:71 – 89, 2014.

[15] K. Kansou, T. Nuttle, K. Farnsworth, and B. Bredeweg. How plants changed the world: Using qualitative reasoning to explain plant macroevolution's effect on the long-term carbon cycle. *Ecological Informatics*, 17:117–142, 2013.

[16] R. King, S. Garrett, and G. Coghill. On the use of qualitative reasoning to simulate and identify metabolic pathways. *Bioinformatics*, 21(9):2017–2026, 2005.

[17] B. Kuipers. *Qualitative Reasoning.* MIT Press, 1994.

[18] P. Salles and B. Bredeweg. Modelling population and community dynamics with qualitative reasoning. *Ecological Modelling*, 195(12):114–128, 2006.

[19] L. Trave-Massuyes, L. Ironi, and P. Dague. Mathematical foundations of qualitative reasoning. *AI Magazine*, 24(4):91–106, 2003.

[20] B. Yang, D. M. Willies, and C. E. Wyman. Changes in the enzymatic hydrolysis rate of avicel cellulose with conversion. *Biotechnology and Bioengineering*, 94(6):1122–1128, 2006.