

*Mechanistic modelling of cancer: some
reflections from software engineering and
philosophy of science*

Cañete-Valdeón, José M. and Wieringa,
Roel and Smallbone, Kieran

2012

MIMS EPrint: **2012.115**

Manchester Institute for Mathematical Sciences
School of Mathematics

The University of Manchester

Reports available from: <http://eprints.maths.manchester.ac.uk/>

And by contacting: The MIMS Secretary
School of Mathematics
The University of Manchester
Manchester, M13 9PL, UK

ISSN 1749-9097

Mechanistic modelling of cancer: some reflections from software engineering and philosophy of science

José M. Cañete-Valdeón · Roel Wieringa · Kieran Smallbone

Received: 17 August 2012 / Revised: 26 October 2012 / Accepted: 29 October 2012 / Published online: 13 November 2012
© Springer-Verlag Berlin Heidelberg 2012

Abstract There is a growing interest in mathematical mechanistic modelling as a promising strategy for understanding tumour progression. This approach is accompanied by a methodological change of making research, in which models help to actively generate hypotheses instead of waiting for general principles to become apparent once sufficient data are accumulated. This paper applies recent research from philosophy of science to uncover three important problems of mechanistic modelling which may compromise its mainstream application, namely: the dilemma of formal and informal descriptions, the need to express degrees of confidence and the need of an argumentation framework. We report experience and research on similar problems from software engineering and provide evidence that the solutions adopted there can be transferred to the biological domain. We hope this paper can provoke new opportunities for further and profitable interdisciplinary research in the field.

Keywords Cancer · Systems biology · Mechanistic models · Knowledge representation · Explanation

Communicated by: Sven Thatje

J. M. Cañete-Valdeón (✉)
School of Informatics, University of Sevilla,
Sevilla, Spain
e-mail: jmcv@us.es

R. Wieringa
Computer Science Department, University of Twente,
Enschede, The Netherlands

K. Smallbone
Manchester Centre for Integrative Systems Biology,
University of Manchester,
Manchester, UK

Introduction

Mechanistic modelling has been employed in biology at least since the past century: A paradigmatic example is the Hodgkin–Huxley giant squid model for action potentials (Hodgkin and Huxley 1952), regarded a stunning accomplishment for which the authors shared the Nobel Prize in 1963. In the field of cancer research, mechanistic models are focused on describing specific aspects of tumour progression in order to explain the underlying biological processes which drive them (Anderson and Quaranta 2008). An early, highly cited example is Fearon and Vogelstein's work on colorectal tumorigenesis (Fearon and Vogelstein 1990). During the twenty-first century, the number of mechanistic models of cancer has been increasing, in most cases described with mathematics—see Komarova (2005) and Araujo and McElwain (2004) for reviews.

In parallel to advances in biological modelling, philosophy of science has experienced a renewed interest on the issues of mechanistic representation and explanation, whose ideas are elaborations of the rich body of research on modelling developed by philosophers of science during the last part of the twentieth century.

The goal of this paper is twofold. Our first aim is to identify and analyse, with the help of recent philosophical results, some open problems of mechanistic modelling which are undermining its successful transfer to mainstream application by clinical oncologists and tumour biologists. Although they will be detailed throughout the paper, a brief introduction of the identified problems will help to clarify the context of our discussion. To begin with, it is not clear how mathematics can coexist with natural language (still the notation preferred by many biologists) as instruments for describing mechanistic models. This makes the advantages of formal modelling inaccessible to a large majority of the biological community. Second, papers where models are

presented contain scientific hypotheses which are necessary to relate the models with some subject of interest in the world. Hypotheses are complex objects in their own, with varying degrees of confidence. Unfortunately, domain-specific notations such as SBML (Hucka et al. 2010), CellML (Lloyd et al. 2004), and Biocharts (Kugler et al. 2009) are not expressive enough to adequately describe the richness of scientific hypotheses, while the use of mathematics in biology is more focused towards describing the models. Last, cancer researchers often employ unstructured natural language to explain *why* their models are claimed to represent reality. Existing argumentation frameworks might greatly help them to organize, share and validate their reasoning.

All of these issues have been the subject of intense research in software engineering. This discipline deals, on the one hand, with hardware and software constructs which have been engineered in such a way that their properties are largely predictable. On the other hand, software engineering also deals with those parts of the physical and human world that form the so-called “problem world” of the software system. While those parts are in general better understood than the growth of cancer cells, those, at least, that involve human interaction are far less well understood (Jackson 2001). It is therefore not far-fetched to hope for useful adaptation of certain software engineering principles to biological systems. The second goal of this paper is to describe some approaches that have worked well in software engineering and to provide evidence on how they can be transferred to the biological domain.

The rest of this paper is organised as follows. The section on “[Related work](#)” introduces related research and concepts from philosophy of science on mechanistic representation. Sections on “[The dilemma of formal and informal descriptions](#)”, “[The need to express degrees of confidence about knowledge](#)” and “[The need of an argumentation framework](#)” present the identified problems, solutions from software engineering and evidence of its application to modelling of cancer. The paper closes with conclusions and references.

Running example: Throughout this paper, we will employ the model of Gatenby and Gillies (2004) for cell–environment interactions in carcinogenesis, described by the authors with natural language and diagrams. The model is mechanistic because it is organised as a process where component parts (such as cells, substrates and basement membrane) interact and produce an emergent behaviour. The process explains that, in pre-malignant lesions, an *adaptation* to intermittent hypoxia happens by persistent metabolism of glucose to lactate (even in aerobic conditions), resulting in increased glucose consumption. Upregulation of glycolysis, in turn, leads to microenvironmental acidosis, which requires an *evolution* towards phenotypes resistant to acid-induced cell toxicity. Evolved cells with upregulated

glycolysis and acid resistance have a powerful growth advantage, which promotes unconstrained proliferation and invasion. We will also consider a mathematical formalisation by Smallbone et al. (2007) of the model of Gatenby–Gillies. The formalisation is a hybrid approach (Anderson 2005), which represents cells in a two-dimensional array governed by rules that determine the discrete state of every cell, together with equations for the continuous metabolite distributions of oxygen, glucose and H^+ .

Related work

Representation in biology has long been a subject of interest in philosophy of science. Waddington (1968), (von Bertalanffy 1968) and others followed a logical empiricist programme. They claimed that biology should seek general theories similar to those found in physics, such as Newton’s theory of gravitation and its elaboration in the *Principia Mathematica*. They regarded scientific theories as linguistic entities, where a small number of fundamental definitions and mathematical axioms constitute the essence of the theory, and axioms are then elaborated deductively in the form of theorems.

For a generation now, a number of philosophers of science have been developing an alternative to the logical empiricist account of scientific theories. Such an alternative is sometimes called the “semantic view of theories”, by way of contrast to the “syntactic” character of theories in logical empiricism (also called the “received view”). Giere (1999, p.122) introduces the term “model-based view” to generically refer to the different “semantic” accounts. This term aims to reflect the common agreement among philosophers that predicates present in theories (linguistic entities such as “pendulum” in classical mechanics) do not refer to anything in the real world. Instead, predicates refer to conceptual, idealised entities called “models”. For example, the predicate “simple pendulum” refers to a mass swinging from a massless string attached to a frictionless pivot, subject to a uniform gravitational force, and in an environment with no resistance. Clearly, this is an ideal object: No real pendulum exactly satisfies any of these conditions, so no real pendulum is a simple pendulum as characterised in classical mechanics. On a model-based view, there is no simple answer to the question “what is the structure of scientific theories?” (Giere 1999, p.98). In contrast, there exist different accounts, opposite in occasions, yet all sharing the former idea of models as a common point.

The model-based view has continued its evolution throughout the twenty-first century with a focus on mechanisms and models of mechanisms (Machamer et al. 2000; Bechtel and Abrahamsen 2005; Glennan 2005; Craver 2006; Woodward 2002). The former term is normally used

to refer to something in the real world, while the latter is employed to denote the representation tool. Some definitions for mechanisms are:

- Machamer et al. (2000, p. 3): “Mechanisms are entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions”.
- Bechtel and Abrahamsen (2005, p. 423): “A mechanism is a structure performing a function in virtue of its component parts, component operations, and their organization. The orchestrated functioning of the mechanism is responsible for one or more phenomena”.
- Glennan (2005, p. 445): “A mechanism for a behavior is a complex system that produces that behavior by the interaction of a number of parts, where the interactions between parts can be characterized by direct, invariant, change-relating generalizations”.

Glennan (2005, pp. 443–444) argues that: “Perhaps because of the realist tendencies of the philosophers involved, most of the literature has focused on the properties of mechanisms themselves and has not said much about the relationships between mechanisms and their models or theoretical representations”. Matthewson and Calcott (2011, p. 738) agree in that “it is sometimes difficult to see where the analysis of mechanisms in the world finishes and where (or if) a discussion of their representations begins”.

Glennan (2005) and more recently other authors such as Matthewson and Calcott (2011) and Schaffner (2007) have elaborated on one of the model-based views of theories: constructive realism (Giere 1988). This philosophical programme regards a scientific theory as an heterogeneous structure consisting, on the one hand, of a family of (non-linguistic) models at different levels of abstraction; on the other hand, a set of (linguistic) hypotheses claiming the *similarity* of models with something in the real world (a system or a class of systems). However, as anything can be similar to anything else in a limitless number of ways, a scientific hypothesis must specify the *respects* in which a model is claimed to be similar to something in the world and the *degrees of accuracy* in which the similarity is claimed.

Matthewson and Calcott (2011) agree with Giere (2006) in that scientists construct different representations of mechanisms according to their different purposes. The authors claim that modelling involves two distinct steps: First, the model is described (which delineates the properties that the model has) and second, the model is deployed: Relevant similarities are found between the properties of the model and the target of interest in the world. According to the authors, “an understanding of a model and the relevant similarities between this model and its target enables scientists to describe, predict or explain part of the world” (Matthewson and Calcott 2011, p.741).

Schaffner (2007, p.146) also elaborates on constructive realism and claims that the typical theory in the biomedical sciences is a structure of overlapping interlevel causal temporal prototypical models. The models of such a structure usually constitute a series of idealised prototypical mechanisms and variations that bear family or similarity resemblances to each other. The author concedes that only a few (but important) theories in biology “have a very broad scope and are characterizable in their more simplified forms as a set of ‘laws’ which admit of mathematically precise axiomatization and deductive elaboration”, such as certain formulations of Mendelian genetics.

In the context of this paper, we employ the term “mechanistic model” in Glennan’s sense, with independence of the description language, whether mathematics (e.g. Smallbone et al. 2007), words (e.g. Fearon and Vogelstein 1990; Gatenby and Gillies 2004) and diagrams (e.g. Bos et al. 2009). Moreover, we adopt constructive realism, with the elaborations made by the mentioned authors, as the working view of scientific theories.

The dilemma of formal and informal descriptions

One difficulty with mechanistic modelling of cancer is actually a linguistic matter: Which is the most convenient language for describing cancer models? Some alternatives are equations, informal diagrams, natural language and structured languages (such as SBML, CellML and Biocharts). The difficulty of developing linguistic tools able to describe tumour growth has long been acknowledged in the literature (Brú et al. 2003). Numerous voices have risen in favour of establishing mathematics as the *lingua franca* for cancer research, based on the conviction that cancer is dominated by *non-linear* system dynamics, whose outcomes cannot be simply determined by verbal or linear reasoning alone but can be adequately expressed in terms of ordinary or partial differential equations, or other mathematical constructs such as cellular automata (Gatenby and Maini 2002, 2003; Maini and Gatenby 2006). Proponents of this approach also advocate a methodological change: Mathematical models can help to formulate biological hypotheses, favouring an active way of researching which contrasts to the more traditional approach of expecting that the general principles of complex biological processes become apparent once sufficient data are accumulated.

However, the adoption of mathematics as a universal language for describing cancer dynamics is not exempt of controversy. Weinberg (2007) points out two main problems: (1) Scientists still do not have enough knowledge about biology and biochemistry to create truly useful, predictive models of tumour development; (2) as a consequence of this lack of knowledge, parameters need to be

assumed or arbitrarily fitted to existing data sets to ensure that the predictive powers of mathematical models conform to actual observations. Weinberg's position has been strongly rebutted by the mathematical biology community. Thus, Gatenby and Maini (2003) argue that similar simplifying assumptions are required in most experimental designs and even simple underlying mechanisms may yield highly complex observable behaviours. However, Komarova (2005) unveils another difficulty: "cancer modelling is often very detached from experimental and clinical cancer research" [...] Because of the involved mathematical expositions, journals that publish theoretical work are mostly inaccessible to the wider, biologically oriented community".

Accordingly, cancer researchers face a dilemma: Dynamics of cancer progression is better described with mathematics but, at the same time, the complex formalism of the descriptions and its fruitful manipulation are not easily accepted by a large part of experimental biologists.

A similar question has long been the subject of intense debate and research in software engineering. During the development of a software system, a number of specifications need to be elaborated. Specifications address different aspects of the system-to-be such as its components, behaviour, physical deployment and user requirements. They cover different levels of abstraction, which help engineers to reason about their design before proceeding to write the program code.

There is a myriad of languages for elaborating software specifications, and some of them have mathematical semantics. Yet, many engineers prefer to use informal diagrams and textual descriptions. The fact is that formal specifications are harder to build and to maintain, but they are susceptible of automatic and semi-automatic analysis, proof and simulation. In contrast, informal specifications are easier to understand and require less bookkeeping, but tool support is very limited. This is even more noticeable in cancer models described with mathematical equations, which produce an emergent behaviour not easy to understand but that can be simulated with software tools.

We will mention two approaches to this problem from requirements engineering, an area of software engineering focused on the systematic handling of the requirements of a software system and especially concerned with describing the real world where the software problem is located.

The first proposal, the so-called "two-button" approach, is working well in the practice of requirements engineering (Lamsweerde A van 2004). It consists of keeping *both* formal and informal descriptions of the subject of interest; the former must be used only "when and where needed" to enable formal analysis techniques (Lamsweerde A van 2009, p.583). Table 1 shows an example, extracted from the requirements of a software system that must schedule meetings for people (Lamsweerde A van 2009, pp.9-12).

Table 1 A two-button specification of a goal, extracted from the requirements of a meeting-scheduling software system (Lamsweerde A van 2009, p.591)

Goal (informal specification): Intended participants shall be notified of the meeting date at least 3 weeks before the meeting takes place.

Goal (formal specification): $\forall p$: Participant, m : Meeting

Scheduled(m) \wedge Invitation(p,m) \wedge \neg Notified(p,m) \Rightarrow
 $\diamond_{\leq m.Date-3w}$ Notified(p,m)

The specification consists of an informal part and a formal part, both of which are complementary views that serve different purposes. Linguistically, the informal part is expressed in English, while the formal part is expressed in linear temporal logic. The intuitive semantics of the formal symbols is: \forall (for all), \wedge (and), \neg (not), \Rightarrow (implies) and $\diamond_{\leq d}$ (some time in the future with deadline d). The formal specification establishes a statement that must be satisfied by any arbitrarily chosen Participant and Meeting, namely p and m . The formal specification employs three predicates (properties that can be true at one time and false at another): Scheduled(m) holds if meeting m is scheduled; Invitation(p,m) holds if participant p has received an invitation to meeting m ; and Notified(p,m) holds if participant p has been notified of meeting m . The expression $m.Date-3w$ means "three weeks before the date when meeting m takes place" (Lamsweerde A van 2009)

The requirements specification includes the following system goal in natural language: Intended participants shall be notified of the meeting date at least 3 weeks before the meeting takes place (Lamsweerde A van 2009, p.591). A formal specification for that goal can be expressed in linear temporal logic (LTL) (Manna and Pnueli 1992). To this aim, two entities have been defined, Participant and Meeting, with arbitrarily chosen instances p and m , respectively. There are also three predicates: Scheduled(m), Invitation(p, m), and Notified(p,m), which are intended to express that some property holds in some arbitrarily chosen current state of the instances: Meeting m has been scheduled; participant p has been invited to meeting m , and participant p has been notified of meeting m . The LTL operator $\diamond_{\leq d}$ means "some-time in the future with deadline d ". The expression $m.Date-3w$ means "three weeks before the date when meeting m takes place".

The informal specification is more expressive and easier to understand. In contrast, the formal specification can be handled by a tool to prove certain properties. For example, a model checker (Clarke et al. 2000) might determine whether a given behavioural specification of the system satisfies the previous goal. It is important to note that, in the two-button approach, informal statements are *not* equivalent to their formal counterparts: If they were, informal statements would be formal after all. Rather, both must be regarded as *complementary* views of the subject of interest.

It is important to note that we are not claiming that languages employed for describing software systems are adequate (in general) for representing biological systems. We argue that the underlying idea, i.e. the two-button approach, may make formal models of cancer more appealing

to many experimental biologists who are not comfortable with pure formal reasoning. Equations and their solutions can be linked to paragraphs and diagrams explaining the underlying processes and conclusions.

An example will help to illustrate our point. Smallbone et al. (2007) formalised the model of Gatenby and Gillies (2004) by a hybrid cellular automaton: an $M \times N$ array of automaton elements (with a specific rule-set governing their evolution) together with oxygen, glucose and H^+ fields, each satisfying reaction–diffusion equations. We have extracted a rule from the model and built a two-button specification with natural language and LTL (see Table 2). The term $\phi_a(i, j)$ denotes the amount of ATP produced by the cell located at row i and column j of the array, and a_0 denotes a certain threshold value, assumed 0.1. The LTL operator \diamond means “eventually in the future”.

The specification of Table 2 can be regarded as two complementary views of the same portion of the mechanistic model described in Smallbone et al. (2007). The informal view may be employed for linear reasoning with words, while the LTL formulation may be employed in a more in-depth research, where hypotheses are examined for internal consistency and compatibility with extant data and where predictions are tested by experiments (Gatenby and Maini 2002).

The other proposal for the formal/informal dilemma is suggested by Jackson (2001) in the *Problem Frames* approach to requirements engineering, where the author is concerned with the question of representing the part of the physical and human world in which a software system is intended to work (the so-called “problem context”). Reality is informal in nature but the system must somehow deal with it to bring about some desired effects, i.e. to satisfy the problem requirements. Jackson (2001, p. 163) expresses this point very clearly: (1) any formalisation of the real world is

Table 2 A two-button specification of one of the rules governing the evolution of a carcinoma, according to the model of Gatenby and Gillies (2004)

Model rule (informal specification): If the amount of ATP produced by a cell falls below a critical threshold value, the cell dies.

Model rule (formal specification): $\forall i, j$
 $1 \leq i \leq M \wedge 1 \leq j \leq N \wedge \phi_a(i, j) < a_0 \Rightarrow \diamond (i, j) = 0$

As in Table 1, the specification consists of an informal part (in English) and a formal part (in linear temporal logic), both of which are complementary views that serve different purposes. The informal specification defines an informal model, while the (interpreted) formal specification defines a formal model. The formal part employs an $M \times N$ array of automaton elements (i, j) . Intuitively, each automaton element (i, j) is interpreted as the state of a different cell; in this case, state “dead” is denoted by $(i, j) = 0$. Term $\phi_a(i, j)$ is interpreted as the amount of ATP produced by a cell and a_0 as the level of ATP required for normal cellular maintenance. The intuitive meaning of the \diamond operator is “eventually in the future”. The remaining formal symbols are interpreted as in Table 1 (Gatenby and Gillies 2004; Smallbone et al. 2007)

at best approximate; and (2) we cannot confidently limit the considerations that might affect the domain properties and behaviour we are interested in: There is always much that has not been considered, and some of it may prove decisive. The author provides us with several tools to represent the world. One of them is a classification of the kinds of things (phenomena) that typically appear in the problem context, such as individuals (phenomena that can be named and distinguished) and states (relationships among individuals that can be true at one time and false at another). Another tool for dealing with informal domains is the use of designations, which establish the empirical meaning of a set of ground terms that we can later use in our formal descriptions. For example:

Dead (c) \approx state : Cell c is not alive.

LowATP (c) \approx state : The amount of ATP produced by cell c has fallen below a critical threshold value.

The first designation introduces the formal term Dead (c). Then, after the designation symbol (\approx), it specifies what kind of phenomenon it is: a state phenomenon. Finally, it gives a *recognition rule* by which we can determine whether what we are observing in the world is, or is not, an instance of the designated phenomenon. According to Jackson, when we write a designation, we are introducing a new class of *observations* that we can make of the world and naming the newly observable phenomena so that we can refer to them in our description (Jackson 2001, pp.163–164). Another class of observations is introduced by the term LowATP (c): the empirical fact that the amount of ATP produced by a cell has fallen below a critical value.

The reader may find Jackson’s designations essentially identical to *correspondence rules* of logical empiricism (see Giere 1988, p.25, for a review). Correspondence rules are used to empirically interpret the non-logical terms of a scientific theory, which is understood by logical empiricists as a formal, logical system. The symbol R in a theory, for example, could be interpreted as standing for the path of a light ray. However, according to Giere (1988), scientific theories make no distinction between the interpretation of “observable” terms and the interpretation of “non-observable” terms. Postpositivist philosophers of science have been unanimous in rejecting correspondence rules in all their manifestations (Giere 1988).

As we argued in “Related work”, this paper does not follow a logical empiricist account of scientific theories, so we do not consider a direct relationship between the statements of a theory (formal or not) and the real world. In constructive realism, such a relationship is indirect through the intermediary of a *theoretical model*. Therefore, we will regard a designation as the linking of a formal term with some concept employed in a model. Thus, Dead (c) is interpreted as “conceptual cell c is not alive” in the model that is being defined. A related problem is that of identification: the linking of a term (or a concept) with

some feature of a specific system (or class of systems) in the real world (Giere 1988). In our example, conceptual cell c in the model might be identified as a specific cell in a certain tumour. We will address this problem in the section on “The need to express degrees of confidence about knowledge” through theoretical hypotheses, which establish a relationship between a model and some system or class of systems in the real world.

Terms introduced by designations can be employed in formal descriptions of mechanistic models. For example, the previous terms can be used to formally state that, if the ATP production of a cell falls below a critical threshold value, the cell will be eventually dead:

$$\text{LowATP}(c) \Rightarrow \diamond \text{Dead}(c)$$

Recognition rules introduce an ancillary problem as it may be extremely difficult to write a rule that holds in the general case (i.e. a universal rule). This is not really important in software engineering because rules only need to refer to the problem context at hand (Jackson 2001, p.165). Focusing on the local context (experiments, clinical evidence) may be also useful when writing recognition rules in cancer models.

A related problem is what Jackson calls the identities concern (Jackson 2001, p.257). The author defines a multiplex domain as consisting of multiple instances of a class of things that are not connected into any structure that identifies them and that do not identify themselves. Multiplex domains are typical in biological systems (e.g. the cells of

a tissue). This is not a problem, unless we need to distinguish one individual from another, as it happens with cells in the model of Gatenby–Gillies.

One solution is to establish a naming convention which maps individuals to identifiers. Another is to introduce a structure that identifies them. The latter has been the approach implicitly adopted by Smallbone et al. (2007) in their formalisation of the model of Gatenby–Gillies: They (imaginarily) arranged cells in an array so that each one is uniquely identified by its relative position as a row and a column.

It is possible to combine designations with the two-button approach. Once a designation has introduced a formal term, it can be (formally) related to some expression of a mathematical model. For example, consider that we are identifying cells by their position in an imaginary two-dimensional array; then it is possible to write terms such as $\text{Dead}(1,1)$ and $\text{LowATP}(1,2)$. Next, we relate the formal terms of the designations with expressions from the mathematical model:

$$\text{Dead}(1,j) \Leftrightarrow (i,j) = 0$$

$$\text{LowATP}(i,j) \Leftrightarrow \phi_a(i,j) < a_0$$

Table 3 shows the rewritten two-button specification of the model rule. Now the formal description is easier to understand and to empirically validate.

Mathematical mechanistic models of cancer described in the literature do not usually pay attention to these issues and assume them implicitly. This is one of the reasons why this kind of models is not as accessible to the wider, biologically oriented community as it should.

Table 3 The formal part of the two-button specification in Table 2 has been extended with designations, the identities concern and relations between designations and mathematical expressions

Model rule (informal specification): If the amount of ATP produced by a cell falls below a critical threshold value, the cell dies.

Model rule (formal specification): $\forall i, j$

$$1 \leq i \leq M \wedge 1 \leq j \leq N \wedge \text{LowATP}(i, j) \Rightarrow \diamond \text{Dead}(i, j)$$

Designations:

- $\text{Dead}(c) \approx \text{state}$: cell c is not alive.
- $\text{LowATP}(c) \approx \text{state}$: the amount of ATP produced by cell c has fallen below a critical threshold value.

Identities concern:

A cell c is identified by its relative position in an $M \times N$ array as row i and column j .

Relation of designated terms to mathematical expressions:

- $\text{Dead}(i, j) \Leftrightarrow (i, j) = 0$
- $\text{LowATP}(i, j) \Leftrightarrow \phi_a(i, j) < a_0$

Designations are employed to interpret certain terms contained in formal statements. In this case, $\text{Dead}(c)$ is interpreted as “conceptual cell c is not alive”, and $\text{LowATP}(c)$ is interpreted as “the amount of ATP produced by conceptual cell c has fallen below a critical threshold value”. The specification makes the *identities concern* explicit for this particular case: while the informal model does not need to individually refer to its (conceptual) cells, the formal model needs to distinguish individual cells by name. However, cells form a *multiplex domain* (i.e. they are not connected into any structure that identifies them, and they do not identify themselves). The formal model solves the identities concern by using a two-dimensional array to identify conceptual cells: they are arranged in the array, and each one is named by its relative position (i, j) . Last, the formal specification relates designated terms with formal expressions. Thus, $\text{Dead}(i, j)$ is formally equivalent to $(i, j) = 0$, and $\text{LowATP}(i, j)$ is formally equivalent to $\phi_a(i, j) < a_0$ (Gatenby and Gillies 2004; Smallbone et al. 2007)

The need to express degrees of confidence about knowledge

The second problem with mechanistic modelling of cancer is related with scientific hypotheses, which, under a semantic view of scientific theories, are necessary to link models with some part of interest in the real world, thus making claims about reality. Contrary to models, hypotheses are linguistic entities and, as such, they may exhibit varying degrees of confidence. Thus, it is a fact that multiple degrees of certainty can be usually found in a single description of a biological process. For example, consider the following claims in the paper by Gatenby and Gillies (2004) (our italics): “it can be *reasonably* surmised that virtually all invasive cancers avidly trap FdG” (p.892); “early carcinogenesis and development of the malignant phenotype *actually* occur in an avascular environment” (p.893); and “these [oxic–hypoxic] temporal cycles are *probably* due to a range of physiological mechanisms” (p.894). Reasonably, actually, and probably are degrees of belief with intuitive, non-mathematical semantics. They are an important part of the scientific hypotheses; further evidence may strengthen or weaken them. However, these degrees are lost in the mathematical formulations of the mechanistic models. Moreover, modelling languages popular in systems biology, such as SBML, CellML and Biocharts, lack constructs for expressing them.

This problem is related to the question of how to understand languages employed in software engineering to represent the physical world, an issue that we addressed in previous research (Cañete-Valdeón 2008). In that paper, we concluded that constructive realism is more appropriate to understand modelling languages than the hard realism advocated by many software engineers. In other words, for the purpose of representing reality during software development, it is more appropriate to understand modelling languages as instruments for defining conceptual entities than as tools for expressing literal truths about the real world. Constructive realism has another advantage: It explicitly considers uncertainty in hypotheses by specifying a range of similarity degrees between certain respects of the model and some subject of interest in the real world. This idea can be applied to mechanistic models in biology. For example, consider the following paragraph from the model of Gatenby and Gillies (2004, p. 894):

“Oxic–hypoxic cycles in tumours have been measured to occur with periodicities of minutes, hours, or days [...] These temporal cycles are probably due to a range of physiological mechanisms. Relatively rapid oxic–anoxic cycles can occur because of fluctuations in haematocrit and vasomotion. Variations occurring over days probably involve vascular remodelling or

cycles of neoangiogenesis and regression due to hypoxia-induced expression of secreted vascular endothelial growth factor [...]”.

In the paragraph, the authors make a claim about the observed oxic–hypoxic cycles in tumours, which constitute the part of reality under study. Under a constructive realist view, what the authors are describing is an idealised process where fluctuations in haematocrit and vasomotion produce rapid oxic–anoxic cycles and where vascular remodelling and other processes produce the longer cycles. That idealised process is a mechanistic model, and, according to constructive realism, it may be more or less *similar* to reality. A hypothesis is then necessary to explicitly specify how much. In the quoted paragraph, the hypothesis is: “These temporal cycles are probably due to a range of physiological mechanisms”. The hypothesis points out a model respect: the causes of the oxic–hypoxic cycles, and qualifies its similarity to the world with a degree: “probably”. As another example, the statement:

$$\text{LowATP}(i,j) \Rightarrow \diamond \text{Dead}(i,j)$$

is trivially true for any conceptual cell in the model presented in Smallbone et al. (2007), as it is characterising something that is nothing but an idealised process. However, in the real world, will a cell *always* die if the state LowATP holds? Or is it rather *likely*? Or is it just *possible*? If we want to employ a model for representing reality, we must make a claim in the form of a scientific hypothesis. For example:

The condition $\text{LowATP}(i,j) \Rightarrow \diamond \text{Dead}(i,j)$ is *likely/possibly/inevitably* to happen in real tumours.

Here the model respect is indicated by $\text{LowATP}(i,j) \Rightarrow \diamond \text{Dead}(i,j)$, and the choices for the (informal) degree of accuracy are “likely” or “possibly” or “inevitably”. The hypothesis makes a claim about real tumours (the identified class of systems in the world). Of course, a hypothesis may be true or false, and arguments and empirical evidence can be presented in favour or against.

So far, we have uncovered the complexity of scientific hypotheses in papers containing mechanistic models, and we have stressed the importance of explicitly distinguishing them from the models themselves. However, how can hypotheses be related with their corresponding models? This question has only recently been addressed in software engineering. In this discipline, modelling languages offer constructs for describing models but not for writing hypotheses. To this aim, in Cañete-Valdeón (2008), we proposed writing hypotheses as *annotations* to the models. Annotations are a construct typically present in software engineering modelling languages for extending their expressivity in ways which were not originally conceived when the languages were developed. The semantics of

annotations are outside the language semantics. Typically software tools are built to capture the annotations attached to a model and then interpret them in a tool-dependent way.

This approach can also be applied to modelling languages of biology. The SBML Level 3 Core Specification includes the *Note* and *Annotation* constructs for storing information intended to be seen by humans or to be processed by software tools, respectively. In the case of CellML, there is a proposal (currently under discussion) for annotating models (Cooling 2011). As for Biocharts, it is based on statecharts (Harel 1987), a software engineering language which can be visually annotated.

To illustrate this point, consider the model of Fearon–Vogelstein (1990), which establishes a series of stages in the development of a colorectal tumour. The authors depict a state diagram whose transitions are driven by genetic alterations. We have reconstructed the main stages of the model using Biocharts, a language that combines statecharts (which capture the high-level state-based strata of system behaviour) with an appropriately well-defined language (preferably a diagrammatic one) for specifying the lower-level dynamics of pathways and networks. Figure 1 shows a statechart for the process described by Fearon and Vogelstein (1990). States are represented with rounded boxes, and

state transitions are denoted with arrows. Events may trigger transitions: When an event occurs at the current state, a state change happens. Note that the authors did not specify trigger events for a couple of transitions. The initial state is depicted as a black circle, and the final state is denoted as a porthole.

Regarding scientific hypotheses, Fearon and Vogelstein (1990, p.763) state the following: “Although the [genetic] alterations *usually* occur at characteristic phases of tumor progression, [...] it is the progressive accumulation of changes, rather than their order of occurrence with respect to one another, that is *likely* to be the most important in colorectal tumor progression” (our italics). Therefore, although the authors seem to be confident in the order of events for most tumours, they regard it more likely that *any* order of the events is a better representation of reality than a fixed order. Figure 1 shows these two theoretical hypotheses as annotations of the model. An annotation is represented as a note linked to the statechart through a dashed line. The first hypothesis considers the respect “the order of the events in the statechart” and relates it to the part of reality under study (the sequence of genetic alterations in colorectal tumours) through Degree 1 (“usually equal”). The second hypothesis considers the respect “any permutation of the events in the statechart”, while Degree 2 is “likely

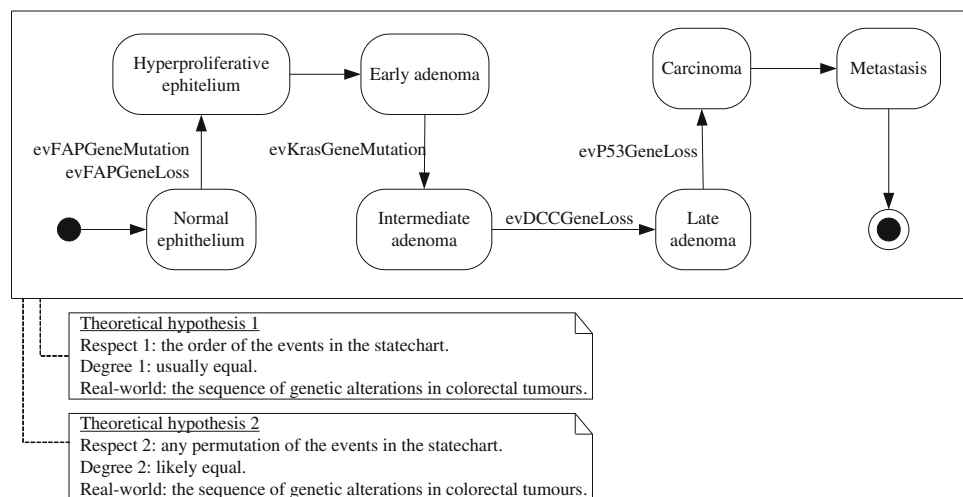


Fig. 1 An annotated statechart (Harel 1987), based on the model of Fearon–Vogelstein for colorectal tumorigenesis (Fearon and Vogelstein 1990). States are denoted with rounded boxes and transitions between states are shown as arrows. The initial state is denoted as a black circle and the final state is depicted as a porthole. The genetic alterations involving oncogenes and tumour suppressor genes have been represented as different events; some examples are: evFAPGeneLoss (meaning the loss of the familial adenomatous polyposis (FAP) gene on chromosome 5q), evFAPGeneMutation (meaning a mutation of the FAP gene) and evKrasMutation (meaning a mutation of the *k-ras* gene on chromosome 12p). The intuitive semantics is: When an event *e* occurs at the current state, the transition labelled with *e* is taken, and the state targeted by the transition is assumed to be reached. Note that two transitions are not labelled in the original reference (Fearon and

Vogelstein 1990). Statecharts are very useful for describing mechanistic models based on discrete processes; however, they are not expressive enough for describing the theoretical hypotheses associated to the models. To overcome this difficulty, we have included two hypotheses as two notes linked to the statechart through dashed lines. A hypothesis asserts that some respect of the model is similar to some identified subject in the physical world to a certain degree. The first hypothesis claims that the order of the events in the statechart is *usually* equal to the sequence of genetic alterations in real colorectal tumours. The second hypothesis claims that the progressive accumulation of alterations, rather than their order of occurrence with respect to one another, is *likely* what happens in real colorectal tumours (Harel 1987; Fearon and Vogelstein 1990)

equal”. Note that it can be established that Degree 2 denotes a belief stronger than Degree 1.

The need of an argumentation framework

The scholarly study of argumentation has been especially rich and fruitful in the legal domain: Tens of argumentation schemes have been identified, analysed and catalogued (Walton et al. 2008). In medicine, however, as in many other fields, there is no explicit meta-analysis of the structure of the arguments supporting the statements contained in a particular piece of research. This constitutes a potential problem when mechanistic models are being employed.

On the one hand, whether formal or informal, a mechanistic model can be regarded as a *designed object* intended to represent reality (in some respects

and to some degrees of accuracy). Important design decisions may need to be justified, such as why certain assumptions were considered and why certain simplifications were applied. In this sense, biology may greatly benefit from a rich body of research in software engineering named design rationale which focuses in capturing the history by which one arrived at a design decision, alternate decisions or the parameters that went into making the decision. To this aim, several languages such as the Decision Representation Language or DRL (Lee and Lai 1991) have been proposed, which may be useful for describing the rationale behind assumptions, simplifications and other design decisions of mechanistic models.

On the other hand, a mechanistic model describes a biological process, and, as such, it may produce a complex, emergent behaviour. The mere statement of

Table 4 A structured argument for one of the hypotheses contained in the model of Gatenby and Gillies (2004)

Argument

Conclusion C: Pre-malignant lesions (such as a polyp or carcinoma) will develop hypoxic regions near the oxygen diffusion limit, provided that premises G1, G2 and G3 are satisfied.

Qualifier: inevitably

Grounds

- Ground G1: Persistent proliferation of the epithelial layer is taking place (some epithelial cells are hyperplastic).
- Ground G2: Blood vessels are physically separated from the epithelial cells by an intact basement membrane (i.e. not breached by an invasive cell).
- Ground G3: In general, oxygen concentrations decrease with distance from a capillary: Oxygenated cells are limited to a distance of less than 150 μm from a blood vessel.

Warrant

Warrant scheme: argument from gradualism (Walton et al. 2008, p.339)

Warrant body

Premise 1: Propositions G1, G2 and G3 are acceptable to the respondent.

Premise 2: the following conditionals are acceptable to the respondent:

- From G1, it follows that (B1) persistent proliferation leads to a thickening of the epithelial layer.
- From G2, it follows that (B2) substrates (such as oxygen and glucose) must diffuse from the vessels across the basement membrane and through layers of (tumour) cells, where they are metabolized.
- From B1 and B2, it follows that (B3) the thickening of the epithelial layer pushes cells ever more distant from their blood supply, increasing the oxygen diffusion distance.
- From B3 and G3, it follows that (B4) epithelial cells at 150 μm from the blood vessels will receive very little oxygen.
- From B4, it follows that (C) hypoxic regions will develop near the oxygen diffusion limit.

Premise 3: The conditional “If G1 and G2 and G3, then C” is not, by itself, acceptable to the respondent, nor are shorter sequences from the grounds to the conclusion (other than the one specified in Premise 2) acceptable to the respondent.

Conclusion: Therefore, proposition C is acceptable to the respondent.

Backing: The process of oxygen diffusion and consumption was modelled by Krogh (1919). Empirical studies by Thomlinson and Gray (1955) showed that viable tumour cells were not observed at distances greater than 160 μm from blood vessels, consistent with Krogh's calculations.

The argument body follows Toulmin's scheme. The conclusion, qualified with a degree of strength, is the hypothesis to be argued. The three grounds are the assumed premises. The warrant is a key component of the argument: It must be convincing enough that the conclusion is true, given the grounds. As warrants have no particular structure in Toulmin's model, we have employed another scheme called “argument from gradualism” (Walton et al. 2008, p.339), which proposes arriving at a conclusion by a sequence of chained conditionals, starting from the grounds and advancing until the conclusion is reached. Argument from gradualism is typically applied in the context of a dialogue between a proponent and a respondent. It consists of three premises and one conclusion. Finally, the last part of the argument is the backing, which provides further evidence for the warrant; in this case, it refers to earlier results by other researchers (Gatenby and Gillies 2004; Smallbone et al. 2007; Walton et al. 2008; Krogh 1919; Thomlinson and Gray 1955)

the emergent properties will, in general, not be trivially accepted by a reader. To this aim, if the model is described formally, a proof of those properties is typically provided, possibly with the help of some software tool such as an equation solver, a theorem prover or a model checker. However, if the model is described with words and informal diagrams, properties of the model need to be informally argued. In any case, it is also necessary to argue scientific hypotheses, understood as statements claiming that some (previously justified) model properties are similar to some subject of interest in the real world to a certain degree (typically an informal qualifier such as *probably*, as explained in the previous section). Arguments for hypotheses include (but are not limited to) references to empirical evidence. In the rest of this section, we will focus on informal arguments.

There are several proposals for structuring informal arguments (Toulmin 1958; Farley and Freeman 1995; Reed and Walton 2003). Some of these have been formalised to a certain extent in artificial intelligence. As an example, we have reconstructed an argument from the model of Gatenby and Gillies (2004) using Toulmin's scheme (see Table 4). This scheme has been successfully employed in software engineering for reasoning on the security of software systems (Haley et al. 2005). In his scheme, Toulmin (1958) considers the *grounds* or premises to arrive at a *conclusion*. A *warrant* is a general argument that acts as a bridge between the grounds and the conclusion. The warrant itself can be further supported by providing a *backing*, which may make a reference to a previous argument or may provide additional pieces of evidence. A *qualifier* expresses the strength with which we are claiming the conclusion.

However, Toulmin does not specify any particular structure for the warrants. To this aim, we have applied an argumentation scheme known as “argument from gradualism” (Walton et al. 2008, p.339) consisting of a chain of inferences which result in the conclusion. In Table 4, we have instantiated the argumentation scheme and have constructed the chain of inferences using properties of the mechanistic model of Gatenby and Gillies (2004).

Toulmin's scheme can accommodate several elements that have appeared throughout this paper: Assumptions can be made explicit as grounds (see, for example, G2 in Table 4); hypotheses can be expressed using the conclusions and their qualifiers (for the degrees of similarity); evidence can be provided in the backing, and the main body of the argument can be stored in the warrant.

The use of structured arguments has the obvious advantage that it makes it easier to validate the underlying reasoning. Komarova (2005) denounces that “everyone knows that most of the [mathematical] models will never be checked”. Then she adds: “Probably the best results can be obtained if theory goes hand in hand with experimental

studies”. Structured arguments can act as a bridge between mechanistic models and hypotheses, on the one hand, and experimental studies on the other.

Conclusions

Mechanistic modelling is a challenging task that compounds the already complex enterprise of cancer research. We agree with Anderson and Quaranta (2008) in the huge benefits of a multidisciplinary approach to the problem. However, in the authors' experience, multi-disciplinary research is not only a scientific question: To be successful, all parties are required hard work in explaining and great effort in understanding. This paper has offered a novel perspective from software engineering and philosophy of science, in a language that we hope can be accessible to a wide community of biological researchers. Our unconventional viewpoint has allowed discovering three problems related with representing cancer: the dilemma of formal and informal descriptions, the need to express degrees of confidence and the need of an argumentation framework. We have offered solutions that have worked well in software engineering, together with examples of their application in tumour research. While these ideas require additional elaboration, we hope to have seeded the motivation for further and profitable interdisciplinary research in the field.

Acknowledgements The authors thank the Editor-in-Chief and the anonymous referees for their useful suggestions. Special thanks are due to Prof. Michael Jackson for his comments and encouragement regarding this work.

References

- Anderson ARA (2005) A hybrid mathematical model of solid tumour invasion: the importance of cell adhesion. *Math Med Biol* 22:163–186
- Anderson ARA, Quaranta V (2008) Integrative mathematical oncology. *Nat Rev Cancer* 8:227–233
- Araujo RP, McElwain DLS (2004) A history of the study of solid tumour growth: the contribution of mathematical modelling. *Bull Math Biol* 66:1039–1091
- Bechtel W, Abrahamsen A (2005) Explanation: a mechanist alternative. *Stud Hist Philos Biol Biomedical Sci* 36(2):421–441
- Bos PD, Zhang XHF, Nadal C, Shu W, Gomis RR, Nguyen DX, Minn AJ, van de Vijver MJ, Gerald WL, Foekens JA, Massagué J (2009) Genes that mediate breast cancer metastasis to the brain. *Nature* 459:1005–1009
- Brú A, Albertos S, Subiza JL, López García-Asenjo J, Brú I (2003) The universal dynamics of tumor growth. *Biophys J* 85:2948–2961
- Cañete-Valdeón JM (2008) On the interpretation of mathematical entities in the formalisation of programming and modelling languages. *Math Struct Comp Sci* 18(6):1017–1030
- Clarke E, Grumberg O, Peled DA (2000) *Model checking*. MIT Press, Cambridge, MA

- Cooling MT (2011) CellML biological annotation metadata specification 2.0. Draft for metadata specification. The CellML Project. <http://www.cellml.org/specifications/metadata/mcdraft/annotationspec>
- Craver CF (2006) When mechanistic models explain. *Synthese* 153(3):355–376
- Farley A, Freeman K (1995) Towards formalizing dialectical argumentation. In: *Proceedings of the 3rd International Conference of the Society for the Study of Argumentation* 156–165
- Fearon ER, Vogelstein B (1990) A genetic model for colorectal tumorigenesis. *Cell* 61:759–767
- Gatenby RA, Gillies RJ (2004) Why do cancers have high aerobic glycolysis? *Nat Rev Cancer* 4:891–899
- Gatenby RA, Maini PK (2002) Modelling a new angle on understanding cancer. *Nature* 420:462
- Gatenby RA, Maini PK (2003) Cancer summed up. *Nature* 421:321
- Giere RN (1988) *Explaining science: a cognitive approach*. University of Chicago Press, Chicago, IL
- Giere RN (1999) *Science without laws*. University of Chicago Press, Chicago, IL
- Giere RN (2006) *Scientific perspectivism*. University of Chicago Press, Chicago, IL
- Glennan SS (2005) Modeling mechanisms. *Stud Hist Philos Biol Biomedical Sci* 36(2):443–464
- Haley CB, Moffett JD, Laney R, Nuseibeh B (2005) *Proceedings of the 3rd Symposium on Requirements Engineering for Information Security*. In: *Arguing security: validating security requirements using structured argumentation*
- Harel D (1987) Statecharts: a visual formalism for complex systems. *Sci Com Program* 8:231–274
- Hodgkin AL, Huxley AF (1952) A quantitative description of membrane current and its application to conduction and excitation in nerve. *J Physiol* 117:500–544
- Hucka M, Bergmann FT, Hoops S, Keating SM, Sahle S, Schaff JC, Smith LP, Wilkinson DJ (2010) The systems biology markup language (SBML). *Nature Precedings*. doi:10.1038/npre.2010.4959.1
- Jackson MA (2001) *Problem frames: analyzing and structuring software development problems*. Addison-Wesley, Boston, MA
- Komarova N (2005) Mathematical modeling of tumorigenesis: mission possible. *Curr Opin Oncol* 17:39–43
- Krogh A (1919) The number and distribution of capillaries in muscles with calculations of the oxygen pressure head necessary for supplying the tissue. *J Physiol* 52:409–415
- Kugler H, Larjo A, Harel D (2009) Biocharts: a visual formalism for complex biological systems. *J R Soc Interface* 7(48):1015–1024
- Lamsweerde A van (2004) Goal-oriented requirements engineering: a roundtrip from research to practice. Keynote talk at the International Conference on Requirements Engineering RE'04. Kyoto, September 10
- Lamsweerde A van (2009) *Requirements engineering: from system goals to UML models to systems specifications*. Wiley, Chichester, UK
- Lee J, Lai KY (1991) What's in design rationale? *Human-Com Interaction—Special Issue on Design Rationale* 6:3–4
- Lloyd CM, Halstead MDB, Nielsen PF (2004) CellML. *Prog Biophys Mol Biol* 85:433–450
- Machamer P, Darden L, Craver (2000) Thinking about mechanisms. *Philos Sci* 67:1–25
- Maini PK, Gatenby RA (2006) Some mathematical modelling challenges and approaches in cancer. In: Nagl, S. (ed) *Cancer bioinformatics: from therapy design to treatment*. 95–107, Wiley, Chichester, UK
- Manna Z, Pnueli A (1992) *The temporal logic of reactive and concurrent systems*. Springer-Verlag, New York, NY
- Matthewson J, Calcott B (2011) Mechanistic models of population-level phenomena. *Biol Philos* 26:737–756
- Reed C, Walton D (2003) Argumentation schemes in argument-as-process and argument-as-product. In: *Proceedings of the conference celebrating Informal Logic @25*, Windsor, ON
- Schaffner KF (2007) Theories models, and equations in systems biology. In: FC Boogerd, FJ Bruggeman, J-HS Hofmeyr and H.V. Westerhoff (eds). Elsevier, Amsterdam
- Smallbone K, Gatenby RA, Gillies RJ, Maini PK, Gavaghan DJ (2007) Metabolic changes during carcinogenesis: potential impact on invasiveness. *J Theor Biol* 244:703–713
- Thomlinson RH, Gray LH (1955) The histological structure of some human lung cancers and the possible implications for radiotherapy. *Br J Cancer* 9:539–549
- Toulmin S (1958) *The uses of argument*. Cambridge University Press, Cambridge, NY
- von Bertalanffy L (1968) *General system theory*. Braziller, New York
- Waddington CH (1968) Introduction. In: *Towards a theoretical biology*. In: Waddington CH (ed) vol. 1. Chicago, IL: Aldine Press
- Walton D, Reed C, Macagno F 2008 *Argumentation schemes*. Cambridge University Press, New York, NY
- Weinberg RA (2007) Using maths to tackle cancer. *Nature* 449:978–981
- Woodward J (2002) What is a mechanism? A counterfactual account. *Philos Sci* 69(3):S366–S377