



# Learning from multi-level behaviours in agent-based simulations: a Systems Biology application

C-C Chen<sup>1\*</sup> and DR Hardoon<sup>2\*</sup>

<sup>1</sup>University College London, UK; and <sup>2</sup>Institute for Infocomm Research, Singapore

This paper presents a novel approach towards showing how specific emergent multi-level behaviours in agent-based simulations (ABSs) can be quantified and used as the basis for inferring predictive models. First, we first show how behaviours at different levels can be specified and detected in a simulation using the complex event formalism. We then apply partial least squares regression to frequencies of these behaviours to infer models predicting the global behaviour of the system from lower-level behaviours. By comparing the mean predictive errors of models learned from different subsets of behavioural frequencies, we are also able to determine the relative importance of different types of behaviour and different resolutions. These methods are applied to ABSs of a novel agent-based model of cancer in the colonic crypt, with tumorigenesis as the global behaviour we wish to predict.

*Journal of Simulation* advance online publication, 5 February 2010; doi:10.1057/jos.2009.30

**Keywords:** behaviour; learning; regression; simulation; systems; system dynamics

## 1. Introduction

Agent-based modelling and simulation (ABMS) is used in Complexity Science and decentralised Systems Engineering to study the relationship between lower-level behaviours between system components and higher-level properties or behaviours, which ‘emerge’ from these. The agent-based model (ABM) itself consists of specifications for different types of agent, each of which represents a particular class or ‘species’. For each type, a set of behavioural rules is defined, which specify how an agent should behave depending on (i) its current state and/or past state(s); and (ii) the input it receives from its environment (usually made up of other agents). In a simulation of an ABM, agent instances exist in a common environment and are able to interact with one another, with each agent behaving according to the set of behavioural rules defined by its type. It is from the individual behaviours and interactions between agents that higher-level properties and behaviours are generated; this is termed ‘emergence’ (Crutchfield, 1994; Holland, 2000; Bedau, 2003; Deguet *et al.*, 2006; Boschetti and Gray, 2007).

However, these higher-level properties and behaviours can also constrain the behaviour of the base entities (‘top-down causation’), and this can give rise to further higher-level properties, and so on. In such cases, the system’s development over time is restricted to a particular set of trajectories (‘self-

organisation’). Furthermore, constraining properties can emerge at many different levels and interact with one another, so there is no straightforward fixed hierarchy. Instead, the system should be treated as a set of dynamic, interacting hierarchies (Rasmussen *et al.*, 2001) or heterarchies (Gunji and Kamiura, 2004; Gunji *et al.*, 2008). The mutual constraints that properties at different levels exert on each other make such systems difficult to analyse and predict.

In this paper, we first show how behaviours at different levels can be specified and detected in a simulation using the complex event formalism. We then apply partial least squares (PLS) regression to frequencies of these behaviours to infer models predicting the global behaviour of the system. To validate this approach, we compare the mean predictive error (MPE) rates obtained from models learned using real data with those learned using randomised data. We also compare the predictive errors of models from different subsets of the behavioural frequencies. To the knowledge of the authors this is the first study done towards showing how specific emergent multi-level behaviours in agent-based simulations (ABSs) can be quantified and used as the basis for inferring predictive models. Our work is based on simulations of a novel ABM of cancer in the colonic crypt, with tumorigenesis as the global behaviour of interest.

The paper will be structured as follows. Section 2 shows how behaviours at different levels can be specified using complex event types (CETs) and then applies this to our ABM of tumorigenesis in the colonic crypt. Section 3 then briefly introduces PLS regression and shows how it can be applied to complex event data to infer predictive models from behaviours at different levels. Section 4 first uses MPE of real and randomised data sets to validate the inferred

\*Correspondence: C-C Chen, Intelligent Systems Group, Department of Computer Science, University College London, Gower Street, London, WC1E 6BT, UK.

E-mail: C.Chen@cs.ucl.ac.uk

DR Hardoon, Institute for Infocomm Research, 1 Fusionopolis Way, # 21-01 Connexis, Singapore 138632.

E-mail: drhardoon@i2r.a-star.edu.sg

models. We then examine differences in MPE between models learned from different data sets representing different sets of behaviours and different temporal resolutions. Section 5 summarises and concludes the paper.

## 2. An inter-level model of tumorigenesis in the colonic crypt

In the context of ABMS, we define an inter-level model to be a model that relates the base level behaviours in an ABM to behaviours defined at different ‘observational’ levels. In order to specify an inter-level model, we require

1. an ABM with a set of state transition rules (STRs) representing behavioural rules or ‘laws’ governing the behaviour and interactions between the modelled real-world entities;
2. a set of explicitly specified higher-level behaviours, which are defined in terms of both the ABM STRs and observed state transitions, as described below; and optionally
3. a set of defined inter-dependency relationships, for example causal, modular between the behaviours at different levels.

The analyses presented in this paper require the first two of these and can be used as a first step to determining whether the third would be plausible (see Section 4.3).

### 2.1. Specifying multi-level emergent behaviours using CETs

In an ABS, certain organisational properties and macro-behaviours can ‘emerge’, which have not been explicitly specified in the agent rules. The complex event formalism introduced below allows us to specify such behaviours as types of ‘observation’ in simulations and is based on two categories of emergence theory:

- Information theoretic interpretations (eg Crutchfield, 1994), which formalise the fact that emergent properties are those that result from viewing the system at different resolutions (distinguishable states) and scopes (components) (see Ryan (2007) for a more detailed account of scope and resolution). For example, a sequence of events can be statistically significantly related to another sequence of events (scope). Usually, only a subset of the events in the first sequence are related to a subset of the events in the sequence; these can be seen to form a minimal causal structure (resolution)—other events are then (statistically) ‘irrelevant’. Two categories of relationship between CETs can be defined, which correspond to scope and resolution:
  - Part/whole (scope): One complex event type,  $CET_1$ , can be a constituent of another  $CET_2$  so that  $CET_1$  is always observed when  $CET_2$  is observed.

- Supertype/subtype (resolution): One complex event type,  $CET_1$ , can be defined more specifically than another  $CET_2$  so that every event that can be classified as  $CET_1$  must be classifiable as  $CET_2$  but not necessarily every event classified as  $CET_2$  is classifiable as  $CET_1$  that is

$$CET_1 \subseteq CET_2$$

- Emergence theories addressing designed/modelled systems (eg Bonabeau and Dessalles, 1997; Kubik, 2003), which also take into account the fact that some behaviours are explicitly specified while others are not. These latter behaviours arise through interactions between the components and are considered to be emergent. In the current context, this corresponds to the distinction between simple and complex events described below.

We define an event as a state transition at a defined level of abstraction. In an ABS, events result from an agent rule being applied; we call these simple events.

Two simple events  $e_1$  and  $e_2$  are said to be of the same type if (a)  $e_1$  and  $e_2$  result from the same agent rule and (b) the scope of  $e_1$ ’s state transition is identical to the scope of  $e_2$ ’s state transition, that is for every component in which a state change occurs in  $e_1$ , there is a component of the same type in which the same type of state change occurs in  $e_2$ .

A complex event,  $CE$ , is defined as either a simple event  $SE$  or two complex events linked by  $\bowtie$ :

$$CE :: SE | CE_1 \bowtie CE_2$$

$\bowtie$  denotes a set of constraints that  $CE_2$  satisfies in relation to  $CE_1$ , for example occurs ‘in the same component’, ‘at the same time’, ‘within distance  $x$ ’. Conceptually, complex events can be thought of as a configuration of simple events in the system space (‘space’ is meant in the general sense and includes all the dimensions represented in the system such as time, physical space, identity).

The type of a complex event is determined both by the types of its constituent events and the relations that hold between them (see Definition 1 and Definition 2). This can be represented as a coloured hypergraph, in which the coloured nodes stand for event types and coloured edges stand for the different relationship types (constraints) existing between pairs of events (Chen *et al*, 2007). Since both events and relation constraints can be defined at different levels, different hypergraphs can be drawn for the same complex event (instance). This reflects the fact that the same event can exemplify more than one type, depending on the level of abstraction.

#### Definition 1

Complex event type (recursively defined). A complex event type is either a simple event type (SET) or two complex events types  $CET_1$  and  $CET_2$  satisfying a particular spatial relation,

a relation type  $\bowtie$  with respect to each other such that the location of  $CET_1$  entails the location of  $CET_2$  and vice:

$$ce :: SET | CET_1 \bowtie CET_2$$

$$\bowtie \rightarrow (l_{CET1} \rightarrow l_{CET2}) \wedge (l_{CET2} \rightarrow l_{CET1})$$

### Definition 2

Complex event type. Two complex events  $ce_1$  and  $ce_2$  are said to be of the same type if, for the hypergraphs representing them  $H_{ce1} = (X_{ce1}, E_{ce1})$  and  $H_{ce2} = (X_{ce2}, E_{ce2})$ :

1. every member of  $X_{ce1}$  has exactly one member in  $X_{ce2}$  that is of the same type and vice versa; and
2. every member of  $E_{ce1}$  has exactly one member in  $E_{ce2}$  that is of the same type and vice versa.

### 2.2. ABM STRs and SETs

The ABM on which our analyses are based aims to model tumorigenesis in the colonic crypt as a function of Adenomatous Polyposis Coli (APC) mutation rate, with only one agent type representing the crypt cells. Due to space constraints, the full set of STRs is not given here, but the STRs reflect the following biological observations when the APC mutation is present:

1. The cell always divides symmetrically Morrison and Kimble (2006).

2. Fitness increases, with a greater effect lower down in the crypt (due to greater levels of the surviving protein) (Boman *et al*, 2004); the effect is also greater when both alleles are mutated.
3. Migration time increases, that is cells move slower so they are more likely to accumulate (Lamprecht and Lipkin, 2002; Aoki and Taketo, 2007); the effect is greater when both alleles are mutated.
4. cMyc is activated (if not already).
5. If cMyc is activated, migration time increases at a greater rate (Krobath *et al*, 2007) (this is modelled by making it equivalent to the rate when both APC alleles are mutated).
6. If Wnt is activated, polyp formation is stimulated (this is modelled by allowing the cell to accumulate, that is does not have to compete with other cell(s) occupying a location) (Fodde *et al*, 1994; Oshima *et al*, 1995; Sansom *et al*, 2004; Andreu *et al*, 2005).
7. There is an increased probability of Wnt activation. This is based on the observation that APC mutated cells behave as if the Wnt-signalling pathway is constantly stimulated (Giles *et al*, 2003; Ilyas, 2005).

Table 1 shows the SETs that are used for specifying CETs representing the higher-level behaviours of interest (see Table 2 and Table 3).

**Table 1** Table showing simple event types (SETs) associated different state transition rules and their biological significance (the biological behaviour represented). Note that not all the SETs of the ABM are shown, only those found in the specified CETs

Simple event type	Biological significance
Asymmetric division ( <i>ASD</i> )	Asymmetric cell division.
Symmetric division ( <i>SD</i> )	Symmetric cell division.
InsertNew ( <i>IN</i> )	A daughter cell is inserted at a particular location in the crypt.
Migrate ( <i>MG</i> )	Cell migrates upwards in the crypt.
MutateAPC1 ( <i>MAPC1</i> )	One allele of APC is mutated.
Activate WNT spatial ( <i>SACTW NT</i> )	Wnt signalling activated by spatial signals.
APC activate WNT ( <i>APCACTW NT</i> )	Wnt signalling activated due to APC mutation.
Compete ( <i>C</i> )	Competition between a pair of cells.

**Table 2** Table defining complex event types for mechanisms associated with APC mutation contributing to tumorigenesis

Complex event type	Specification in terms of simple event types or subtypes
MD	Either (subtype) MSD or (subtype) MAD
MSD	MAPC1 <[sameCell]SD
MAD	MAPC1 <[sameCell]AD
MWD	Either: (subtype) MWDS or (subtype) MWDA
MWDS	(MAPC1 [sameCell, stem]APCACTWNT) <[sameCell]SD
MWDA	(MAPC1 [sameCell, stem]APCACTWNT) <[sameCell]AD
MSWD	Either: (subtype) MSWDS or (subtype) MSWDA
MSWDS	MAPC1 <[sameCell]SACTWNT <[sameCell, stem, mutated]SD
MSWDA	MAPC1 <[sameCell]SACTWNT <[sameCell, stem, mutated]AD

**Table 3** Table defining complex event types for clonal interaction mechanisms contributing to tumorigenesis

Complex event type	Specification in terms of simple event types or subtypes
CC	C[sameClone, differentCell]
CCWIN	Either (subtype) CCMIG or (subtype) CCINS
CCLOSE	C [sameClone, differentCell]IN or C [sameClone, differentCell]MG
CCINS	C [sameClone, differentCell]IN
CCMIG	C [sameClone, differentCell]MG

### 2.3. CETs for the mechanisms underlying APC mutation-driven tumour development

Although the relationship between APC mutation rate and tumorigenesis is linear, there is more than one mechanism underlying this relationship. Each of these mechanisms is effectively a behavioural ‘motif’, which can be formulated as a CET. We can distinguish between two categories of mechanisms believed to contribute to tumorigenesis:

1. Mechanisms directly associated with the APC mutation, which arise as a result of the altered behaviour of individual mutated cells, as defined in Table 2; and
2. Clonal interaction mechanisms, which are by definition higher-level characterisations of cell agent behaviour. CETs representing such mechanisms are defined in Table 3.

In each simulation, different pathways can be active to different degrees. To track each of them, we measure the frequency of events corresponding to different simple and CETs.

### 3. PLS regression analysis of complex event frequencies

Partial Least Squares (PLS) (Geladi and Kowalski, 1985) is a method for constructing a predictive model when the relationships between variables are complex or ill-understood; for example, some may be collinear, some may be non-linearly related. Rather than trying to understand the underlying relationships between variables, however, the main purpose of PLS is to construct a model that is able to predict a set of outcomes (responses), given a set of input variables (factors). In our studies, we use PLS to construct models that are able to predict the degree at which a system-level behaviour (tumorigenesis) occurs, given occurrences of lower-level behaviours (CET occurrence frequencies). This is summarised in Figure 1.

PLS works by projecting to a latent structure. Latent variables (the underlying factors that account for most of the variation),  $X$  and  $Y$ , are extracted from the factors  $F$  (in this case the CET frequencies) and the responses  $R$  (in this case the higher-level behaviour), respectively.  $X$  is then used to predict  $Y$ , and then the predicted  $Y$  is used to construct predictions for  $R$ .

Least squares is the procedure of finding the best fitting curve to data, such that the error of the sum of the squares of the points offset from the curve is minimised. In the process of the optimisation of least squares regression we seek a vector  $\mathbf{w}$  such that it solves

$$\min_{\mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$$

where  $\mathbf{X}$  contains as rows the feature vectors of the samples and  $\mathbf{y}$  contains the outputs.

We are able to consider a more general multivariate regression by taking  $\mathbf{w}$  and  $\mathbf{y}$  to be matrices and the norm to be the Frobenius norm

$$\min_{\mathbf{w}} \|\mathbf{X}\mathbf{W} - \mathbf{Y}\|_F^2$$

Frobenius norm

$$\min_{\mathbf{w}} \|\mathbf{X}\mathbf{W} - \mathbf{Y}\|_F^2$$

Principal Component Analysis (PCA) (Pearson, 1901) is the process of examining the direction of maximum variance within the data. We seek linear combinations of the data that preserves the characteristics of the data while finding directions with maximum variance.

The PCA can be solved by the following optimisation problem

$$\begin{aligned} \max_{\mathbf{w}} \quad & \mathbf{w}'\mathbf{X}'\mathbf{X}\mathbf{w} \\ \text{subject to} \quad & \|\mathbf{w}\| = 1 \end{aligned}$$

where  $\|\mathbf{w}\|$  is the first eigenvector and  $\mathbf{X}$  is centred (ie the origin is moved to the centre of the mass of  $\mathbf{X}$ ). We consider the usage of the features returned from PCA. Using the first  $k$  eigenvectors of  $\mathbf{X}'\mathbf{X}$  as our features and leaving the outputs  $\mathbf{Y}$  unchanged. This translates into two stages; performing PCA and regressing in the feature space given by the first  $k$  principal directions and then minimising the least square error between the projected data and the response. This is also known as Principal Component Regression (PCR). Let  $\mathbf{X} = \mathbf{V}\mathbf{\Sigma}'\mathbf{U}'$  be the Singular Value Decomposition (SVD)<sup>1</sup> of  $\mathbf{X}$ , therefore the data matrix is now represented as  $\mathbf{X}\mathbf{U}_k$  where  $\mathbf{U}_k$  contains the first  $k$  columns of  $\mathbf{U}$ . We describe the

<sup>1</sup>The Singular Value Decomposition (SVD) is a widely used technique to decompose a matrix into several component matrices, exposing properties of the original matrix. Using the SVD, we can determine the rank of matrix, quantify the sensitivity of a linear system to numerical error, or obtain an optimal lower-rank approximation to the matrix.

least squares solution, which will be used in our experiments. In the following we obtain the least squares regression problem

$$\min_{\mathbf{B}} \|\mathbf{X}\mathbf{U}_k\mathbf{B} - \mathbf{Y}\|_F^2 = \min_{\mathbf{B}} \|\mathbf{V}'\Sigma'\mathbf{U}'\mathbf{U}_k\mathbf{B} - \mathbf{Y}\|_F^2$$

where we are able to multiply by an orthogonal matrix  $\mathbf{V}'$  as this does not effect the norm, giving

$$\min_{\mathbf{B}} \|\mathbf{V}'\mathbf{V}\Sigma'\mathbf{U}'\mathbf{U}_k\mathbf{B} - \mathbf{V}'\mathbf{Y}\|_F^2 = \min_{\mathbf{B}} \|\Sigma'_k\mathbf{B} - \mathbf{V}'\mathbf{Y}\|_F^2$$

$\Sigma_k$  is the matrix containing the first  $k$  columns of  $\Sigma$  and similarly let  $\mathbf{V}_k$  contain the first  $k$  columns of  $\mathbf{V}$ . We find that

$$\begin{aligned}\bar{\Sigma}'_k\mathbf{B} &= \mathbf{V}'\mathbf{Y} \\ \mathbf{B} &= \bar{\Sigma}_k^{-1}\mathbf{V}'_k\mathbf{Y}\end{aligned}$$

Where  $\bar{\Sigma}_k^{-1}$  is the symmetric square matrix containing the first  $k$  columns inverse of  $\Sigma_k$ .

Following the SVD of  $\mathbf{X}$  we find that  $\mathbf{V}_k = \mathbf{X}\mathbf{U}_k\bar{\Sigma}_k^{-1}$ , allowing us to express  $\mathbf{B}$  as

$$\mathbf{B} = \bar{\Sigma}_k^{-2}\mathbf{U}'_k\mathbf{X}'\mathbf{Y}$$

Showing that the components are computed as an inner product between the features and the data matrix weighted by the inverse of the eigenvalues, the critical measure of the different coordinates is their covariance with the data matrix  $\mathbf{X}'\mathbf{Y}$  suggesting that rather than seeking directions that give

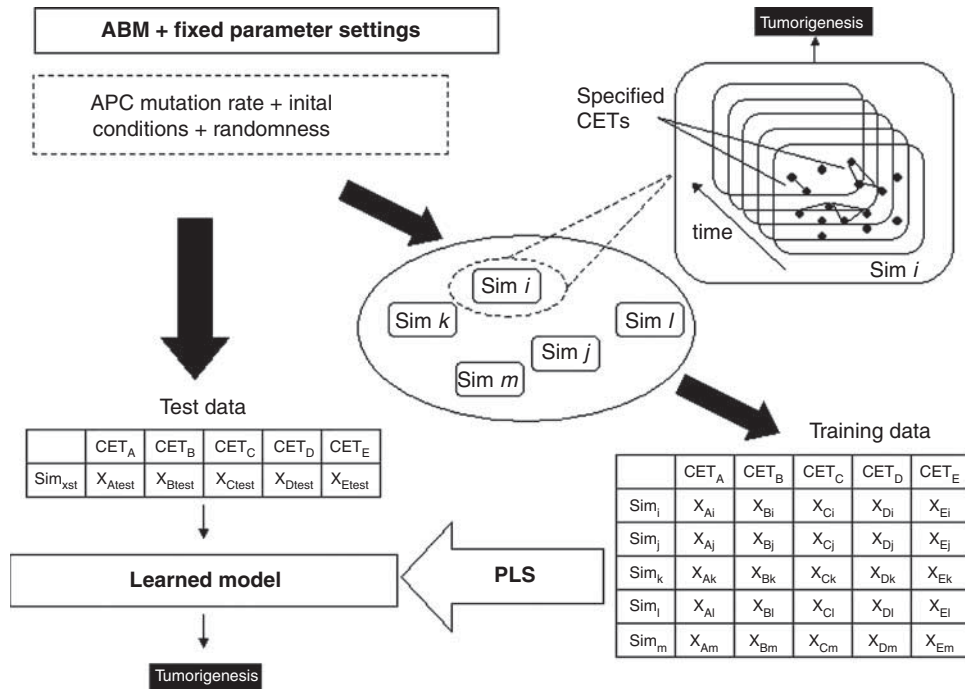
maximum variance we should seek direction that maximise the covariance.

#### 4. Validating and comparing models learned from different data sets

In our studies, models were inferred from the following data sets:

1. *APC and clonal dominance (CD)*: APC mutation rate and initial CD (not *CETs*).
2. *Clonal interaction CETs overall*: The total frequencies of clonal interaction *CETs*, as defined in Table 3.
3. *Mutation-driven CETs overall*: The total frequencies of mutation-driven *CETs*, as defined in Table 2.
4. *CETs overall*: The total frequencies of both clonal interaction *CETs* and mutation-driven *CETs*.
5. *CETs 300ts*: The frequencies of both clonal interaction and mutation *CETs* extracted at nine 300-time step intervals from time step 0 to time step 2100.
6. *SETs 300ts*: The frequencies of the 27 *SETs* (see Table 1) extracted at nine 300-time step intervals from time step 0 to time step 2100.
7. *Both CETs and SETs 300ts*: The frequencies of both the specified *CETs* and the *SETs* extracted at 300-time step intervals, giving 302 independent variables (IVs) in total.

For each data set, 100 different models were learned from 80-simulation subsets of 100 simulations, with the remaining



**Figure 1** Outline of method used to infer predictive models from *CET* frequencies. The colonic crypt ABM and fixed parameter settings generate a set of simulations. Differences in the *CET* frequencies between simulations are due to differences in APC mutation rate, simulation initial conditions, and randomness throughout simulations. *CET* frequencies are used as the input observations for learning the predictive model using PLS. The learned model can then be used to predict tumorigenesis from *CET* frequencies.

20 simulations used as the test sample to test the predictive validity of the learned models. The MPE for each model is the mean of the discrepancies between the values (for the tumorigenesis measures) predicted by the inferred model given the event frequencies of the test sample, and the actual values observed in the test sample.

#### 4.1. Validating models learned from real against randomised data set models

For each 80-simulation subset, we inferred two models: one from real data, and the other from randomised data, and calculated their respective predictive errors. The means of the MPEs across models were then calculated for models learned from each data set. Figure 2 shows that for all data sets, the PLS models learned from real data had lower predictive error rates than those learned from randomised data sets, and the results of  $t$ -tests between real and random data models confirm that in all cases, the difference is significant ( $p=0.001$ , two-tailed).

#### 4.2. Comparing models learned from different data sets

Table 4 shows the MPEs for the models learned from the different model sets (where each model set contains the models learned from a particular data set) and the sizes of the data sets in terms of number of independent variables. Paired  $t$ -tests comparing the MPEs for the different model sets showed differences between every model comparison ( $p=0.001$ , two-tailed). The following observations were made:

- The data set for the clonal interaction  $CETs$  performs better than that for mutation-driven  $CETs$  ( $t=9.365$ ), suggesting that among the specified  $CETs$ , those representing clonal interactions are more dominant than those representing mutation-driven behaviours in determining the degree of tumorigenesis.
- The data set with both mutation-driven  $CETs$  and clonal interaction  $CETs$  ( $t=10.217$  and  $t=3.343$ , respectively)

performs better than either alone, suggesting that the specified mutation-driven  $CETs$  still have significant effects on the degree of tumorigenesis.

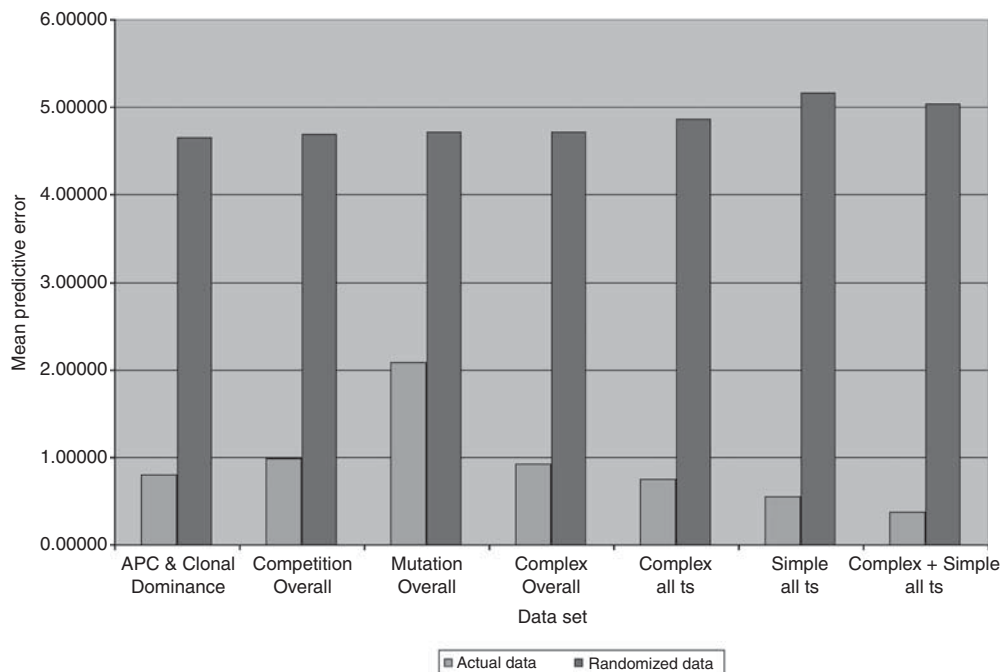
- The data set for  $SETs$  performs better than that for  $CETs$  ( $t=21.093$ ). This is consistent with the idea that the  $SET$  set contains higher resolution information.
- For  $CETs$ , the data set with greater temporal resolution ( $CETs$  300ts *versus*  $CETs$  overall) performs better ( $t=17.255$ ), suggesting that the higher temporal resolution gives us additional information.
- The data set with both  $SETs$  and  $CETs$  performs better than either the  $SET$  set or the  $CET$  set on its own ( $t=21.093$  and  $t=39.695$ , respectively), suggesting that the higher-level behaviours specified in the  $CETs$  give us additional information that is not contained in the  $SETs$  alone.
- The MPE for models learned from APC mutation and initial CD are relatively low. In terms of data efficiency, these models perform best, since only two independent variables are used.

#### 4.3. Interpretation of MPE differences

In the complex systems framework, differences between MPEs can be given a number of different interpretations. If  $DS1$  and  $DS2$  are two model sets containing models learned from two different data sets with different observation types ( $CETs$ ), the fact that  $MPE_{DS1} < MPE_{DS2}$  for phenomenon  $X$  can be given several different interpretations, for example, that the types of input observations of  $DS1$  are better indicators of: (i) the ‘causes’ of  $X$  (which is itself subject to a number of different interpretations; for example Williamson, 2007; Pearl, 2000); (ii) lower-level processes or states underlying  $X$ ; or (iii)  $X$  itself. Validation of one or more explicit models describing the interdependency relations between the observation types (eg causal, modular), such as structural equation models or Bayesian nets would be required to determine the plausibility of different interpretations. However, by establishing the significance of the building blocks of such explicit inter-level models, PLS analysis gives an indication as to whether such explicit models are worth pursuing at all.

**Table 4** Table showing mean predictive errors and standard deviations of the models learned from the different data sets. The size of each data set is also given

Data set	Mean pred. error	Mean pred. error randomised	Mean difference	SD	SD randomised	SD difference	No. IVs
APC and CD	0.80144	4.65264	3.85120	0.12372	0.63137	0.58944	2
Clonal interaction $CETs$ overall	0.98023	4.68417	3.69948	0.21170	0.64342	0.70927	5
Mutationdriven $CETs$ overall	2.08583	4.71405	2.76500	1.00772	1.65530	1.14076	9
$CETs$ overall	0.92283	4.71631	3.81141	0.11180	0.69374	0.71220	14
$CETs$ 300 ts	0.74734	4.87059	4.12325	0.09124	0.69643	0.69782	98
$SETs$ 300 ts	0.54860	5.16293	4.61432	0.07858	0.74636	0.73826	189
$CETs$ 300 $SETs$ all ts	0.37663	5.03825	4.66162	0.05008	0.72304	0.71908	287



**Figure 2** Graph showing the predictive error rates for PLS models learned from different data sets. For each data set, a model was also learned from randomised data in the set. In all cases, the model learned from the randomised data set performed significantly worse, that is the predictive rate was significantly lower.

## 5. Summary and conclusions

In this paper, we have introduced novel methods for analysing ABSs and applied these to a biological ABM of tumorigenesis in the colonic crypt. CETs were specified to represent behaviours at different levels of abstraction and then used as input data for learning predictive models using PLS regression. The learned models could then be used to predict tumorigenesis, a higher-level system behaviour from lower-level behaviours (the specified CETs). Analysis of the MPEs of models learned from different sets of CETs provided a means of determining the relative importance of different types of behaviours and possible interactions between behaviours. This can provide the first step towards explicit modelling of the interdependencies between behaviours at different levels.

*Acknowledgements*—Chih-Chun Chen acknowledges the advisory support of Christopher D. Clack and Sylvia B. Nagl as part of a multi-disciplinary collaboration between the departments of Computer Science and Oncology at UCL. David R. Hardoon acknowledges financial support from the EPSRC project Le Strum, (<http://www.lestrum.org>) EP-D063612-1 and from the EU project PinView, (<http://www.pineview.eu>) FP7-216529.

## References

Andreu P *et al* (2005). Crypt-restricted proliferation and commitment to the paneth cell lineage following APC loss in the mouse intestine. *Development* **132**: 1443–1451.

- Aoki M and Taketo M (2007). Adenomatous polyposis coli (apc): A multi-functional tumor suppressor gene. *J Cell Sci* **120**: 3327–3335.
- Bedau MA (2003). Downward causation and the autonomy of weak emergence. *Principia* **3**: 5–50.
- Boman BM *et al* (2004). Colonic crypt changes during adenoma development in familial adenomatous polyposis. *Am J Pathol* **165**(5): 1489–1498.
- Bonabeau E and Desselles JL (1997). Detection and emergence. *Intellectica* **2**(25): 85–94.
- Boschetti F and Gray R (2007). Emergence and computability. *Emergence: Complexity and Organisation* **9**: 120–130.
- Chen CC, Nagl SB and Clack CD (2007). Specifying, detecting and analysing emergent behaviours in multi-level agent-based simulations. In: Wainer G (ed.) *Proceedings of the Summer Simulation Conference, Agent-directed Simulation SCS*.
- Crutchfield JP (1994). The calculi of emergence: Computation, dynamics, and induction. *Physica D* **75**: 11–54.
- Deguet J, Demazeau Y and Magnin L (2006). Elements about the emergence issue—A survey of emergence definitions. *ComplexUs* **3**: 24–31.
- Fodde R *et al* (1994). A targeted chain-termination mutation in the mouse apc gene results in multiple intestinal tumors. *Proc Natl Acad Sci USA* **91**: 8969–8973.
- Geladi P and Kowalski BR (1985). Partial least-squares regression: A tutorial. *Anal Chim Acta* **185**: 1–17.
- Giles RH, van Es JH and Clevers H (2003). Caught up in a wnt storm: Wnt signaling in cancer. *Biochim Biophys Acta* **1653**(1): 1–24.
- Gunji Y-P and Kamiura M (2004). Observational heterarchy enhancing active coupling. *Physica D* **198**(1–2): 74–105.
- Gunji Y-P, Sasai K and Wakisaka S (2008). Abstract heterarchy: Time/state-scale re-entrant form. *Biosystems* **91**(1): 13–33.

- Holland J (2000). *Emergence—From Chaos to Order*. Oxford University Press: Oxford.
- Ilyas M (2005). Wnt signalling and the mechanistic basis of tumour development. *J Pathol* **205**(2): 130–144.
- Krobath K *et al* (2007). Lack of adenomatous polypis coli protein correlates with a decrease in cell migration and overall changes in microtubule stability. *Mol Biol Cell* **18**(3): 910–918.
- Kubik A (2003). Toward a formalization of emergence. *Artif Life* **9**: 41–66.
- Lamprecht SA and Lipkin M (2002). Migrating colonic crypt epithelial cells: Primary targets for transformation. *Carcinogenesis* **23**(11): 1777–1780.
- Morrison SJ and Kimble J (2006). Asymmetric and symmetric stem-cell divisions in development and cancer. *Nature* **441**(7097): 1068–1074.
- Oshima M *et al* (1995). Loss of apc heterozygosity and abnormal tissue binding in nascent intestinal polyps in mice carrying a truncated apc gene. *Proc Natl Acad Sci USA* **92**: 4482–4486.
- Pearl J (2000). *Causality: Models, Reasoning and Inference*. Cambridge University Press: Cambridge.
- Pearson K (1901). On lines and planes of closest fit to systems of points in space. *Philos Mag* **2**(6): 559–572.
- Rasmussen S *et al* (2001). Ansatz for dynamical hierarchies. *Artif Life* **7**: 329–353.
- Ryan A (2007). Emergence is coupled to scope, not level. *Nonlinear Sci, Complexity* **13**(2): 67–77.
- Sansom OJ *et al* (2004). Loss of apc in vivo immediately perturbs wnt signalling, differentiation, and migration. *Genes Dev* **18**: 1385–1390.
- Williamson J (2007). Causality. In: Gabbay DM and Guenther F (eds). *Handbook of Philosophical Logic*. Springer: Dordrecht, The Netherlands, pp 89–120.

*Received March 2009;  
accepted November 2009 after two revisions*