Genome-wide mouse embryonic stem cell regulatory network self-organisation: a big data CoSMoS computational modelling approach

Richard B. Greaves[1], Sabine Dietmann[2], Austin Smith[2], Susan Stepney[1], and Julianne D. Halley[1]

[1] York Centre for Complex Systems Analysis, University of York, UK
[2] Wellcome Trust-Medical Research Council Cambridge Stem Cell Institute, University of Cambridge, UK

Abstract. The principal barrier to gaining understanding of embryonic stem (ES) cell regulatory networks is their complexity. Reductionist approaches overlook much of the complexity inherent in these networks and treat the ES cell regulatory system as more or less equivalent to the sum of its component parts, studying them in relative isolation. However, as we learn more about regulatory components it becomes increasingly difficult to integrate complex layers of knowledge and to develop more refined understanding. We seek better control of the complexity inherent in non-equilibrium ES cell regulatory networks undergoing lineage specification by developing computer simulations of self-organisation using the CoSMoS approach. Simulation, together with the hypothesis that lineage computation occurs at the edge of chaos, should allow us to investigate the driving of gradual accumulation of network complexity 'from the bottom up'. Here, we present the first step in this design process: use of the CoSMoS approach to develop a highly abstracted model and simulation of regulatory network activity driven by just single pluripotent transcription factors (TF), but at genome-wide scales. We investigate three TFs in isolation: Oct4, Nanog and Sox2, central elements of the core pluripotent network of mouse embryonic stem cells. This provides a suitable basis for future modelling of multiple interacting TFs.

1 Introduction

Mathematical or computational frameworks and tools are indispensable in the study of cell regulatory networks [Bornholdt, 2005; Zandstra and Clarke, 2014] because functions, traits and pathologies are rarely caused by single genes [Hartwell et al., 1999; Weatherall, 2001; Bornholdt, 2005]. However, the principal challenge that prevents comprehensive understanding (and simulation) of regulatory networks is their complexity [Mesarovic et al., 2004]. Indeed, in the era of systems biology, the icon for molecular biology is the 'hairball' graph, which illustrates how everything seems to interact with almost everything else [Ferrell, 2009; Lander, 2010]. High-throughput technologies generate such large volumes of data that there is concern about how to grasp the big picture [Bray, 2003; Howe

et al., 2008; Driscoll, 2009] and most data sets are not being used to their full potential.

Here we present the first iteration of a novel computational framework to interrogate the complexity of stem cell regulatory networks. We employ a previously described theoretical framework based on the notion that the backbone of stem cell fate computation is provided by the critical-like self-organisation of transcription factor (TF) regulatory networks [Halley and Winkler, 2008; Halley et al., 2009, 2012].

We apply the modelling framework CoSMoS [Andrews et al., 2010; Stepney and Andrews, 2015; Stepney et al., 2016], which is specifically designed to capture the emergent properties of complex systems, and to guide the engineering of trustworthy computer simulations, i.e., those that are scientifically valid, useful and credible to third parties.

The models and results in this paper report on the first iteration of the CoSMoS design cycle. Here, we design and calibrate simulations of single TFs in isolation. This single TF version of the full model is not biologically realistic; its purpose is to serve as a building block of complexity that will be iterated in our next work.

The structure of the paper follows the patterns defined in the CoSMoS approach outlined in §2. This progresses through the definition of scope and the model of the scientific domain in §3, then the development of the simulation software in §4, and use of the simulation to run experiments and explore system behaviour in §5. We conclude with some reflections on the process in §6, and discussion of further work in §7.

2 The CoSMoS approach

The CoSMoS approach [Andrews et al., 2010; Stepney and Andrews, 2015; Stepney et al., 2016] enables the construction and exploration of computer simulations for the purposes of scientific research. It describes a series of models and other components that need to be specified, designed, and implemented in order to build and use a fit-for-purpose simulator. The approach is guided by considering the simulator to be a form of *scientific instrument* [Andrews et al., 2012] that needs to be carefully designed, built, calibrated and used in a manner appropriate to the specific research questions.

The CoSMoS approach is encapsulated as a *pattern language* [Alexander et al., 1977]. The CoSMoS patterns provide guidance on what to do at the various stages of a CoSMoS simulation project [Stepney, 2012; Stepney et al., 2016]. We structure this paper explicitly in terms of these patterns.

To guide the reader through the pattern structure, we reproduce in boxed text a brief overview of the pattern: the pattern name and *intent*, a short phrase describing what should be done; and, where applicable, any *components* (including sub-patterns) that can be used to decompose the intent. We use section subheadings to capture the specific pattern names (named with initial capitals,

such as *Research Context*) and other components (named in lower case, such as *success criteria*) and their position in the overall pattern structure.

We start at the top level of a simulation project, which is formed of three phases per iteration of the project.

CoSMoS pattern: *CoSMoS Simulation Project:* Develop a basic fit-for-purpose simulation of the complex scientific domain of interest.

The components of a *CoSMoS Simulation Project* are:
- carry out a *Discovery Phase*
- carry out a *Development Phase*
- carry out an *Explorations Phase*
- iterate as required

In this paper we report on the first iteration of our simulation, comprising a simulation of a single TF branching process. This provides the basis for the next iteration, which will add multiple interacting TFs. The next three sections document the results of carrying out this first iteration of each of these three phase patterns, structured in terms of their sub-patterns.

3 Discovery phase

CoSMoS pattern: *Discovery:* Decide what scientific instrument to build. Establish the scientific basis of the project: identify the domain of interest, model the domain, and shed light on scientific questions.

The components of the *Discovery* phase are:
- identify the *Research Context*
- define the *Domain*
- do *Domain Modelling*
- define the *Expected Behaviours*
- Argue Appropriate Instrument Designed (omitted here)

3.1 Discovery > Research Context

CoSMoS pattern: *Research Context:* Identify the overall scientific context and scope of the simulation-based research being conducted.

The components needed to identify the *Research Context* are:
- provide a thumbnail *overview* of the research context
- document the *research goals* and project scope
- agree the *Simulation Purpose*, including criticality and impact
- identify the *Team* members, including the Domain Scientist, the Domain Modeller and the Simulation Engineer, their roles, and experience
- document *Assumptions* relevant to the research context
- note the available *resources*, timescales, and other constraints
- determine *success criteria*

– decide whether to proceed, or walk away

Discovery > Research Context > overview

The context of this research is the investigation of a conceptual approach: self-organisation at the edge of chaos. We have argued that if the activity of single transcription factors can be described as critical-like branching processes, their interplay should define a critical-like genome-wide interference pattern that captures in some way the nature of the entire pluripotency transcription factor regulatory network [Halley et al., 2012].

Here we build a simulation based on the representation of TFs as *branching processes*. The mathematical concept of a branching process (BP) is as follows. Consider a population of individuals. At time t each individual i produces a next generation of m_i offspring individuals, with the value of m_i drawn from some probability distribution. Let the average number of offspring produced be μ. If $\mu > 1$, then the process is supercritical and the number of individuals grows without bound. If $\mu = 1$ then the system is critical and can either give rise to more individuals in the next step or lead to dissipation of the process. If $\mu < 1$ then the process goes to extinction.

Our model of TF BPs builds on this idea, and also allows the TFs to *interact* in such a way as to cause the regulatory network to self-organise at the edge of chaos. We capture the activity of single TFs as BPs in order to predict the interplay of multiple TFs and the emergent nature of the entire TF regulatory network, hypothesised to operate in a critical-like state [Halley et al., 2012].

For a TF to be stably expressed, its BP must be supercritical [Halley et al., 2012]. Therefore, by modelling the activity of TFs known to be expressed in mouse embryonic stem cells, we link the perturbation of a TF's cistrome (portion of the genome in which the TF displays some activity) with a dynamic and distributed description of TF activity. This is a prerequisite to being able to simulate the entire TF regulatory network of an ES cell, as argued in [Halley et al., 2012]. The TFs called Oct4, Sox2 and Nanog are central elements of the core pluripotent network of mouse embryonic stem cells. In the first instance, the current work will allow us to calibrate our simulation for these three TFs in isolation, that is, to characterise how their associated TFBPs propagate in the absence of interference.

Our iterative approach to the development of the full simulation commences with the simplest possible system: the operation of one transcription factor at genome-wide scales. We will later add layers of further complexity, testing and calibrating as we go.

A model of a single pluripotent TF in isolation is far from complete and is not biologically realistic. It is only when multiple TF BPs are simulated in parallel that we can expect to generate the interference patterns predicted to underpin circuitry self-organisation. As greater numbers of pluripotency TFs are included in the model, we anticipate that our simulations will become increasingly biologically realistic. In future work we will augment the complexity of the

computational model in a stepwise manner, adding detail and refining assumptions as we progress, and increasingly be able to provide insights not accessible by other means.

Discovery > Research Context > research goals

The overall research goals of this work are:

1. to create a simulation of Branching Process Theory (BPT) as applied to embryonic stem cell differentiation
2. to use this simulation to validate the application of BPT in this context
3. to make the simulation available for more general use

Here we report on the first iteration, of a single TF branching process.

Discovery > Research Context > Simulation Purpose

> CoSMoS pattern: *Simulation Purpose:* Agree the purpose for which the simulation is being built and used, within the *Research Context*.
>
> The components of the *Simulation Purpose* are:
> – define the role of the simulation
> – determine the criticality of the simulation results

Simulation role: The role of the simulation is exploratory: to provide evidence of the usefulness of BPT as a model of decision making in stem cell differentiation. The simulation will be used to investigate which values of the average branching ratio are required to set up a sustainable TF branching process.

Simulation criticality: The simulation work is being used to explore the suitability of of a particular approach, BPT, in the domain. The simulation results are not safety, security, or financially critical: they will not be used directly in the development of any products.

Discovery > Research Context > team

The three main CoSMoS roles are fulfilled by the team members in the following way:

- *Domain Scientist*: Halley, an expert on BPT as applied to stem cell differentiation, backed up by a domain expert in ES cell biology (Smith), and a data collection expert (Dietmann)
- *Domain Modeller*: Greaves, with CoSMoS domain modelling experience, backed up by a further CoSMoS modelling expert (Stepney)
- *Simulation Engineer*: Greaves, with agent based simulation engineering experience

Discovery > Research Context > Assumptions

CoSMoS pattern: *Document Assumptions:* Ensure assumptions are explicit and justified, and their consequences are understood.

The components of *Document Assumptions* are:

- identify that an assumption has been made, and record it
- for each assumption, determine its nature and criticality
- for each assumption, document the reason it has been made
- for each reason, document its justification, or flag it as "unjustified" or "unjustifiable"
- for each assumption, document its connotations and consequences
- for each critical assumption, determine the connotations for the scope and fitness-for-purpose of the simulation
- for each critical assumption, achieve consensus on the appropriateness of the assumption, and reflect this in fitness for purpose arguments
- revisit the simulation scope in light of the assumption, as appropriate

A.1 Cistrome data can be provided by processed ChIP-Seq data
reason It is the data we have
justification This is one standard use for ChIP-Seq data
consequence ChIP-Seq data is variable across measurements, so we will need to check the robustness of our results to this variation

A.2 It is sufficient to consider only the key pluripotency transcription factors: Nanog, Oct4, Sox2
reason As a first step in providing insight, we consider the three TFs widely acknowledged to be central components of the core pluripotent network
justification See for example [Boyer et al., 2005]
consequence We will not be able to determine the effect of further TFs. However, it should be straightforward to incorporate further TF data into the multi-cistrome model.

A.3 We can use mouse data as a suitable proxy for data from human ES cells
reason Suitable mouse data is more readily available; mouse ES cells have an unambiguous 'ground state'; so mouse data is a good basis for evaluating the TF BP model
justification Although effective manipulation of human ES cells is a long term goal, here we are only assessing the TF BP model
consequence We cannot extrapolate results to the human system

Discovery > Research Context > resources, timescales, other constraints

The project has a one year duration. The Domain Scientist is employed full time, and Simulation Engineer part time.

The work has access to a local computer cluster, for running simulations and gathering performance metrics.

The team members are split between York (Halley, Greaves, Stepney) and Cambridge (Smith, Dietmann)

Discovery > Research Context > success criteria

1. a single-cistrome simulator that exhibits the expected behaviours, and can be used as the basis for multi-cistrome simulator development
2. a multi-cistrome simulator that can justify the use of the TF BP model to analyse stem cell fates

   This paper documents the first iteration: the single-cistrome simulator

3.2 Discovery > Domain

CoSMoS pattern: *Domain:* Identify the subject of simulation: the real-world biological system, and the relevant information known about it.

The components are:

- draw an explanatory *Cartoon*
- provide an *overview* description of the domain
- provide a *Glossary* of relevant domain-specific terminology
- Document *Assumptions* relevant to the domain
- define the *scope and boundary* of the domain – what is inside and what is outside
- identify relevant *sources*: people, literature, data, models, etc

Discovery > Domain > Cartoon

CoSMoS pattern: *Cartoon:* Sketch an informal overview picture of the *Domain.*

Figure 1 is a cartoon of the regulatory process. A single gene regulation and its expression is conceptually relatively straightforward; the complex interplay of multiple interacting regulatory processes is not.

Discovery > Domain > overview: embryonic stem (ES) cell biology

Modern, high-throughput laboratory techniques routinely provide large-scale datasets including complete genome sequences, dynamic measurements of gene expression, extensive lists of regulatory proteins and RNAs, and *in vivo* occupancy of DNA by TFs, cofactors and nucleosomes [Ay and Arnosti, 2011]. Such datasets facilitate the investigation of ES cell regulatory networks. To create a complete multi-layered model of a stem cell network one should exploit these big data to bridge gaps between the phenotypic behaviour of whole cells and key regulatory molecules [Xu et al., 2010].

We need to capture the results of multiple high-throughput experiments within a logical and transparent conceptual and computational framework in order to facilitate the interrogation of multiple layers of complex regulatory information. Our initial model is based on the complete genome sequence of mouse embryonic stem cells and on ChIP-Seq data that capture the density of

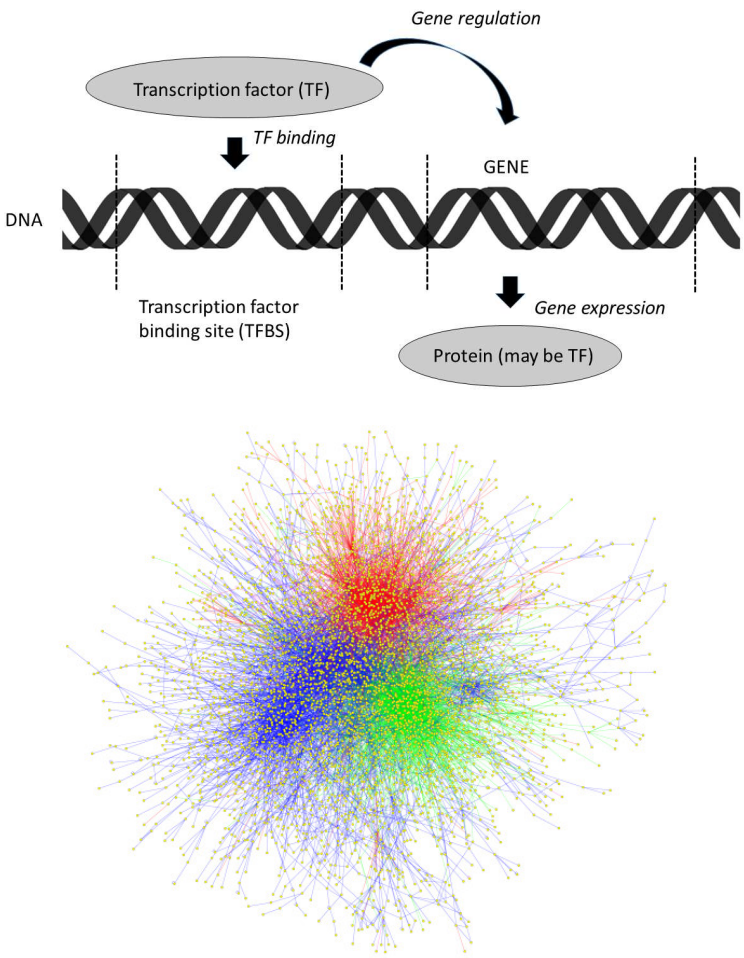

Fig. 1. Domain > Cartoon: (top) The regulatory process: a TF protein binds to DNA at the BS, thereby regulating production of protein (which may be a TF) from the corresponding gene (gene expression). (bottom) Expressed proteins may include other TFs that can regulate expression of other genes: a 'hairball graph' of the human proteome and its binding interactions [Ferrell, 2009, fig.1]

TF binding sites throughout the genome. TFs operate in parallel, influencing each other; according to our hypothesis, they produce genome-wide interference patterns that capture in some way the predicted nature of the entire pluripotent circuitry.

Embryonic stem (ES) cells have the potential to produce all of the different cell types within the body, but this behaviour cannot yet be efficiently exploited *in vitro*. We have considerable knowledge of the component parts of the regulation of ES cells maintained under precise external conditions [Martello and

Smith, 2014], but during normal development many different types of regulatory factors interact, enabling cells to respond flexibly to changing environments. The regulatory network of single ES cells is therefore some function of both cell intrinsic and cell extrinsic variables.

Here we assume that pluripotency is a state of individual ES cells. ES cells exit pluripotency via a transient 'primed' state that facilitates cell fate computation [Nichols and Smith, 2009]. Our knowledge of this exit process and the transient primed state is incomplete, partly because it is difficult to obtain data from transient cell states [Teles et al., 2013]. The process of pluripotency exit itself is intrinsically disorganised and/or chaotic in order for it to integrate intrinsic and extrinsic information and compute cell fate. According to our conceptual framework, regulatory circuitries compute cell fate trajectories via 'critical-like dynamics' at the edge of chaos [Halley et al., 2012].

Nanog, Oct4 and Sox2 form part of the core pluripotency circuitry of ES cells [Boyer et al., 2005]. Oct4 in particular seems central to understanding pluripotency. Oct4 expression level is closely regulated, with deviations either above or below a certain expression range resulting in differentiation [Niwa et al., 2000]. It has been suggested that protein complexes, in which Oct4 is involved, help to establish a dynamic competition between individual elements, serving to buffer the differentiation-promoting activity of Oct4 [Muñoz Descalzo et al., 2013].

Fluctuations are inevitable in any system that has many degrees of freedom. At static equilibrium, such fluctuations ultimately disappear but under non-equilibrium conditions, fluctuations are often great enough to drive reorganisation toward new dynamic states [Nicolis and Prigogine, 1977; Chaisson, 2004]. If continual driving is experienced, complex spatiotemporal patterning usually results and systems are said to have 'self-organised' [Nicolis and Prigogine, 1977; Gollub and Langer, 1999; Ball, 2001].

In biology, the growth and development of organisms occurs far from equilibrium. The stem cell regulatory networks that facilitate these processes are replete with positive and negative feedback loops and nonlinear interactions. When faced with overwhelming complexity, the natural tendency of humans is to either reduce, simplify or ignore it. Reductionist thinking makes systems (a) easier to think about, (b) easier to consider manipulating, and (c) easier to predict, provided non-equilibrium driving is minimal.

Over the last few decades, there has been increasing awareness of the limitations of the reductionist approach [Crutchfield et al., 1986; Farmer and Packard, 1986; Bak and Paczuski, 1993; Parisi, 1993; Kauffman, 1995] and it has become clear that some laws of nature cannot be deduced by resolving more detail [Vicsek, 2002]. This so called 'new era of physics' focuses on developing complex behaviour out of simplicity, instead of the traditional reductionist approach that reduced complexity to its simplest possible form [Kadanoff, 1987; Anderson, 1991; Parisi, 1993]. Non-equilibrium driving can have profound consequences on system behaviour, a realisation that contrasts with our natural tendency to assume systems are near equilibrium or at least show some steady state behaviour. Equilibrium and reductionist thinking pervades most scientific disciplines [Bak

and Paczuski, 1995; Ball, 1999, 2001; Ekeland, 2002], including molecular and stem cell biology.

The differentiation of pluripotent cells in the early embryo is a fascinating non-equilibrium process that results in the production of numerous specialised cell types. More than 600 different proteins have been implicated in exit from a naïve pluripotent state and control of early state transitions in the mouse [Kalkan and Smith, 2014]. As our focus shifts from individual components to complex communication networks, experimental studies have become more difficult. Not only do central features of complex networks, such as robustness, prevent straight forward analysis and interpretation of network behaviours, but many experiments cannot be performed because of ethical reasons surrounding the use of human embryos.

Computer simulation sidesteps the ethical, moral and political issues surrounding use of human embryos. It therefore represents an alternative route to gaining new insight in to this promising field of regenerative medicine. Our overarching aim is to gain sufficient understanding so that any cell type of therapeutic interest can be generated effectively at will.

Discovery > Domain > Glossary: terms and acronyms

> CoSMoS pattern: *Glossary:* Provide a common terminology across the simulation project.

The main biological terms used in the various models are:

binding site (BS) : section of DNA that binds a given TF and influences transcription of associated genes

branching process (BP) : the mathematical model underlying inspiration of the TF BP framework being investigated here

ChIP-Seq : a technique to identify the binding sites of transcription factors on DNA

cistrome : the portion of the genome associated with a specific TF; a pattern of genome-wide binding sites to which the TF displays some activity

pluripotent stem cell : a cell capable of generating all the cell types present in the adult body

segment : the genome data is segmented, into say 10k or 50k base-pair sequences, in order to apply the TFBP framework

transcription factor (TF) : a protein that binds to DNA to influence transcription of the associated gene

Discovery > Domain > assumptions

See §3.1 for the *Assumptions* pattern requirements.

A.4 The genome can be modelled as a set of overlapping TF cistromes without needing epigenetic factors

reason We are looking only at TF segments, and the pluripotent state can be induced by TFs alone

justification See, for example, [Kim et al., 2008]

consequence Behaviours facilitated by other factors, such as epigenetics, will be unseen in the model

A.5 a TFBS is either bound or unbound, there is no partial TF binding

reason not enough data to say otherwise

A.6 a segment can be either activated or deactivated, there are no differing amounts of activation

reason Simplification: the data does say whether a segment has one or more binding sites

justification This is the first iteration; we will revisit the necessity/impact of this assumption in later iterations

consequence We will not be able to separate out behaviours of groups of genes in a segment. In order to do so, we could use smaller segments. But segments cannot be made too small, else we would lose correlations between related TFs.

A.7 we can investigate cell decision making by modelling an individual cell, not a population

reason cells have internal decision making, although they can also be influenced by their environment

justification See, for example, [Loh et al., 2006]

consequence We will not be able to investigate population-level decision making

Discovery > Domain > scope

- single cell model
- single transcription factor model
- later iterations will add more, coupled TFs, and more interacting cells

Discovery > Domain > sources

- Domain scientists
- Biological literature, as referenced in the various overviews
- Chip-seq data for various cistromes (source: Dietmann)

3.3 Discovery > Domain Modelling

CoSMoS pattern: *Domain Modelling:* Produce an explicit description of the relevant domain concepts.

The components of *Domain Modelling* are:

- *collaborate* with the identified Domain Scientist
- draw an explanatory *Cartoon*
- discuss and choose the *Modelling Approach* and level of abstraction
- build the *Domain Model* using the chosen modelling approach
- build the *Data Dictionary*

– document *Assumptions* relevant to the domain model
– Argue Domain Model Appropriate (omitted here)

Discovery > Domain Modelling > collaborate

The lead domain scientist (Halley) and the domain modellers (Greaves, Stepney) collaborated closely throughout the development of the domain model, translating and abstracting the conceptual TF BP model into a form suitable for simulation.

The domain scientists (Halley, Smith, Dietmann) collaborated on refining the research context.

The simulation engineer (Greaves) collaborated with the the data collection expert (Dietmann) on the form and content of the biological data provided.

Discovery > Domain Modelling > Cartoon

See §3.2 for the *Cartoon* pattern.

Due to the structure of our Domain Model description, the *Domain Modelling Cartoon* is presented in the section on the TF BP model (figure 4), and should be read in in that context.

Discovery > Domain Modelling > Modelling Approach

CoSMoS pattern: *Modelling Approach:* Choose an appropriate modelling approach and notation.

A central part of this design process is to develop the simplest possible working model at each stage of the modelling process. This 'agile' approach ensures that simulation code is not unnecessarily complicated. It also helps to ensure that if a coding problem is found, it is simple matter to backtrack to the last working model.

The domain model is captured using UML, in anticipation of an agent-based, object-oriented design and implementation of the simulator.

Discovery > Domain Modelling > Domain Model

Our domain modelling gives rise to several models at different levels of abstraction: a specifically biological stem cell model of regulatory networks, a model simplifying detailed transcription regulatory networks using branching process theory, and a generic abstract model, which we refer to as the 'sparking posts' model.

Note that the sparking posts model could also be used as a domain model for other biological phenomena as captured by branching process theory, such as patterns of information flow in the human brain.

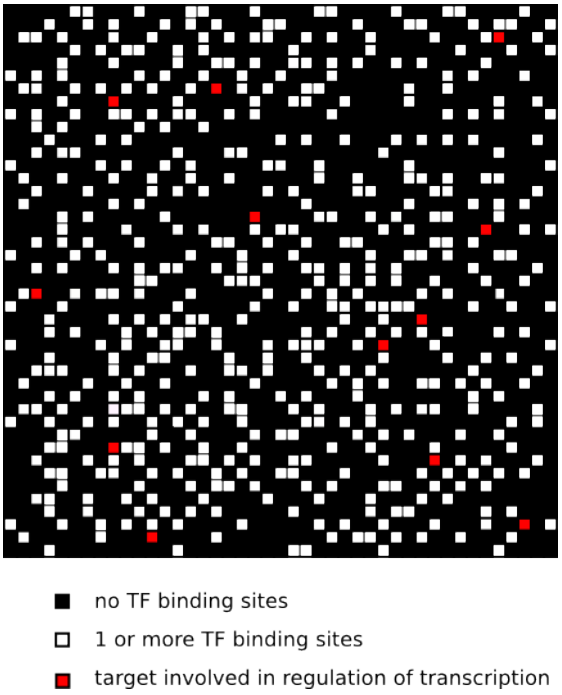

Fig. 2. A representation of a set of ChIP-Seq data for a cistrome (part of the genome relevant to a specific TF). Each square represents a 50kb segment of DNA. A white square is a segment that contains at least one BS site for a product that is not a TF. A red squares is a segment that contains at least one BS site for a product that is a TF. A black square is a segment that does not belong to this cistrome.

Regulatory network

We have mouse genome data including the suite of BSs within it. For convenience and simplicity, we divide this sequence in to 50 kilobase (kb) segments, any of which may or may not contain binding sites for a particular TF of interest. If a 50kb segment contains a binding site for our transcription factor, X, then the segment is said to be part of the X cistrome.

Data about the locations of the transcription factor binding sites, in relation to the gene segments in the model, is provided experimentally by ChIP-Seq data. Figure 2 is a representation of ChIP-Seq data.

The regulatory network components can be captured in a model such as that shown in figure 3. However, we abstract away from many of these 'hairball' inducing details, and consider the system instead in terms of the TF BP model.

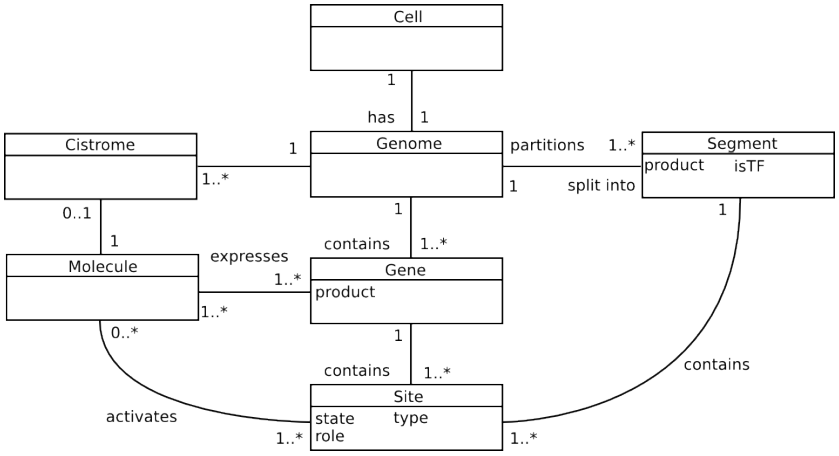

Fig. 3. Class diagram model of a stem cell pluripotency regulatory network. The stem cell has a genome comprised of genes, which can alternatively be described as a cistrome (or set of cistromes), each being comprised of segments of gene which may or may not contain transcription factor binding sites.

Transcription Factor Branching Process model

A common approach to understanding cell regulatory processes is the application of concepts, tools and techniques developed in mathematics, physics or computer science [MacArthur et al., 2008]. Network representations, for example, can accommodate multiple types of data within a single visual illustration that provides an overview of regulatory pathways and components [Gallagher and Appenzeller, 1999; MacArthur et al., 2008]. As already mentioned, empirically-derived interaction networks can be difficult to interpret, often appearing as a 'hairball' graph as regulatory mechanisms are increasingly dissected.

We use here a novel way to visualise and simulate genome-wide regulatory network interactions. Our coarse-grained approach does not require details of binding constants prerequisite for most ODE models of stem cell regulation. In many previous computational or mathematical models of stem cell regulatory networks, TFs are represented as single nodes with binary (on/off) behaviour. Here, we use a different approach that captures TF activity as a dissipative branching process that propagates within the bounds imposed by the TF's unique cistrome.

Unlike reductionist models that capture TF activity using single variables in an equation, in our model we explicitly represent a background delocalisation of TF activity throughout the genome. We can visualise the activity of each TF's BP as a kind of gateway through which regulatory information pertaining to the TF passes over time.

The TF BP model allows a decoupling between details of BS constants and the emergent effect of TF activity throughout the genome. Instead of struggling with countless (often unknown) binding constants, we consider the overall flow

of regulatory information at genome-wide scales. It is thus more suitable for attempts to discover how the ES cell regulatory network behaves as a whole during computation of lineage choice. Through this more coarse-grained methodology, we hope to discover complex interactions that can easily be overlooked by studies that focus on only a handful of key regulatory components at a time.

The potential binding of a TF to target regions throughout the genome is determined by ChIP-Sequencing. The data set or 'footprint' for a given TF comprises a unique pattern of TF-DNA interactions that is somewhat dependent upon the precise methods used to infer interactions. The precise footprint for a specific TF may vary between different experimental datasets. Such 'fuzziness', rather than being a nuisance, is intrinsic to the TF BP model.

If we understand the activity of any given TF as a branching process of regulatory information propagating through time, it makes sense that there will be some correlation between observed TF expression and the saturation of target sites influenced by TF activity. The significance of this important point will become clearer in later work, when we simulate multiple cistrome data sets. Here, we focus on simulating a single TF's BP to introduce the groundwork for our approach.

Figure 4 presents a Cartoon of the TF BP model. Each square in the figure corresponds to a 50kb segment of the mouse genome. Black squares represent segments that contain no BSs for the TF of interest, while red and white squares represent segments with at least one BS for the TF of interest. The difference between a red and white segment lies in their products. A red segment has products that include TFs, whereas none of the products of a white segment is a TF. Henceforth, when we refer to a 'red' segment we mean a gene segment that can bind TF and thus become stimulated into transcribing further TFs.

We capture the countless (ill-defined or unknown) cascades of gene activation via TF production and feedback as a branching process in which TFs produce other TFs while also regulating the remainder of the genome. There are potentially three qualitatively different types of behaviour for any TF_X branching process. Firstly, the cistrome X is saturated and the TF_X gene is continually and stably expressed. Alternatively, there is the opposite type of emergent behaviour, with TF_X expression occurring at a very low noisy level that is not sustainable unless TF_X is supported by continual activation of the TF_X gene via some external signal. Finally there is a dynamic intermediate between these extremes where a branching process only just percolates through the TF_X cistrome. In all cases, the targets of TF_X are divided in to two types: (1) dissipative targets that do not propagate information back in to the TF_X cistrome and (2) amplifying targets that are either TFs themselves and capable of propagating information or code for signalling molecules that are involved in signal transduction.

We define an average branching ratio, called m, for our gene regulation branching process. That is to say that once transcribed, a gene (or gene segment in our case) will produce m product molecules (in this single cistrome model these will all be the TF that binds to binding sites within the cistrome of interest). If the activated site is associated with TF products then new TFs are

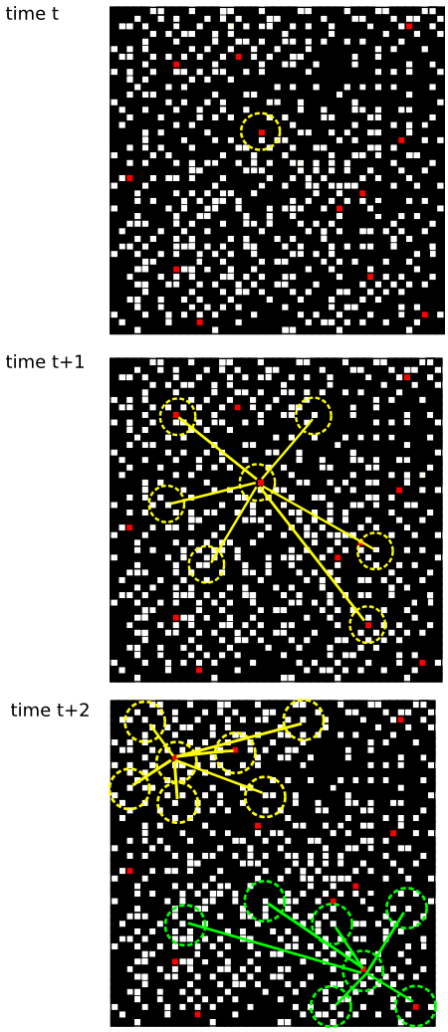

Fig. 4. Domain Modelling > Cartoon: A branching process representation of the overall flow of regulatory information, which serves as the basis of our simulation. At t, assume the circled red segment is activated. At time $t+1$ this will activate m further randomly chosen segments (arrows), and itself deactivate. At time $t+2$, any of these newly activated segments that are themselves red, will each activate a further m randomly chosen segments, and deactivate.

produced and these can bind to other TF binding sites in the system. In this way, up to m segments will be activated in the next time step of the algorithm. In the time step after this each of the activated segments can go on to activate m further segments and so on as illustrated in Figure 4.

This TF BP model is built on the classical BP theory outlined in section 'Domain > overview', and is adapted in the following ways:

- m is related to the BP branching factor μ, but is not the same, because here the m 'offspring' include both white and red segments, yet only red segments go on to produce further 'offspring'.
- In the supercritical case, the number of offspring cannot increase without bound, but only up to the number of relevant segments in the cistrome.
- The individulas are segments, and do not 'die' at the end of a generation; rather they can be reused (reselected) in subsequent generations.

Domain Model: Sparking Posts

In order to model a branching process, we produce our domain model in terms of a metaphor. To capture the nature of critical-like self-organisation hypothesised to underpin lineage computation, we have reduced the system to a 'sparking posts model'. This computational model is used to define the backbone of critical-like self-organisation upon which other layers of complexity are elaborated.

The TF BP representation of our system is modelled as a 'sparking posts' representation of the cistrome in which each segment is modelled as a metal 'post' which emits 'sparks' once it has been activated by an incoming spark emitted by another post in the previous timestep. The sparks represent the TF products of the genes contained within a given segment and are therefore the principal mode of communication between cistromes, the genome being effectively the sum of all cistromes in the system.

So the Domain Model is as follows.

Consider an *arena* containing metal *posts*, some *red*, some *white*. The arena is an abstraction of a particular cistrome; the posts are abstractions of the segments containing BSs (red and white squares in figure 2); red posts are abstractions of segments that express TFs (red squares in figure 2).

Posts may be *active* (on) or not. In a timestep, an active red post emits m *sparks*. A post being active is an abstraction of a gene in a segment being activated; a red post sparking is an abstraction of an activated gene expressing a TF.

Posts become deactivated after they have sparked. A spark lands on a random post in the arena (that is, the model is aspatial), and activates it.

Continued propagation of sparks relies on the activation of sufficient red posts at each timestep.

Figures 5 and 6 capture this Domain Model.

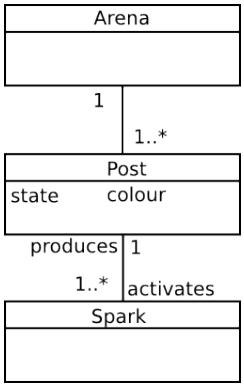

Fig. 5. A class diagram capturing the sparking posts model components. An Arena has multiple Posts; Posts produce Sparks, and are activated by Sparks.

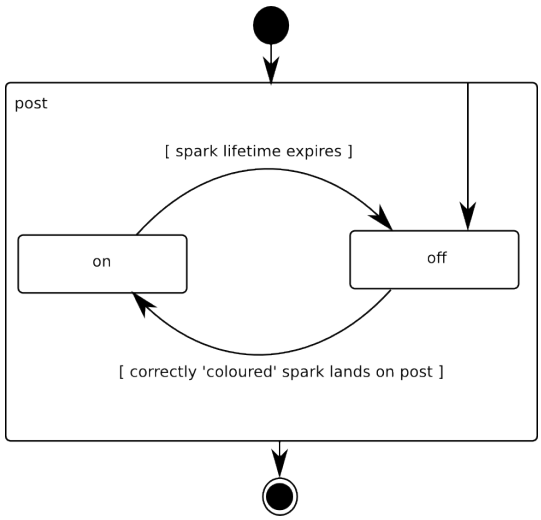

Fig. 6. A state diagram of a post. Posts are initially off; become activated (on) if a spark lands; then become deactivated in the next timestep.

Discovery > Domain Modelling > Data Dictionary

CoSMoS pattern: *Data Dictionary:* Define the modelling data used to build the simulation, and the experimental data that is produced by domain experiments and the corresponding simulation experiments.

The sparking post model's parameters and variables are shown in figure 7. Figure 8 shows the values of some of these parameters for the cistromes of interest here.

p	total number of posts in the arena
r	number of red posts
m	sparks emitted per active red post
s_0	number of red posts active initially
t	timestep
s_t	number of red posts active at timestep t

Fig. 7. Sparking post model: (top) parameters, constant during a simulation run; (bottom) variables, changing during a simulation run

	Nanog	Sox2	Oct4
p	4310	3330	2540
r	631	542	466
r/p	0.146	0.163	0.183
$p/r = m_c$	6.8	6.1	5.5

Fig. 8. The values of the parameters p (number of posts, or segments in the cistrome) and r (the number of red posts, or red segments in the cistrome) for the TFs investigated in this study

Discovery > Domain Modelling > assumptions

See §3.1 for the *Assumptions* pattern requirements.

First, we have some assumptions related to the TF BP model, which we note as they have an impact on the sparking posts model.

A.8 the product of a TF producing segment is the TF whose cistrome we are modelling
reason An assumption underlying use of the TF BP model
justification The TF may not be directly produced; there may be a cascade of production, but the TF BP model collapses this cascade. We are investigating this model.
consequence This is an abstraction from the biology, made to allow us to model the highly complex processes. If it works, this abstraction could also provide an approach to include other features such as epigenetics and mRNAs in a tractable model.

A.9 the identity of the TFs produced during transcription is irrelevant in the single cistrome model
reason An assumption underlying use of the TF BP model
justification The TF BP model assumes that the relevant scale of computation is the cistrome level, abstracted from specific details of the individual TFs

Assumptions directly related to the sparking posts model are:

A.10 a spark from a post can hit any post with equal probability: there is no notion of a 'distance' between posts

reason an aspatial model
justification the TF BP model collapses a potential cascade of TFs into a single 'proxy' TF. This cascade would lose any spatial dependence in the DNA.

A.11 a post cannot be hit by more than one spark per timestep: there is no notion of different 'capacity' posts
reason follows from assumption A.6

Discovery > Domain Modelling > Expected Behaviours

CoSMoS pattern: *Expected Behaviours:* Describe the expected emergent behaviours of the underlying system.

The 'sparking posts' domain model forms the basis for subsequent simulation development.

We can form a much simpler version of the model, in order to help understand the effect of noise. Since there are a finite number of posts, stochastic fluctuations will occur, and sparks might occasionally miss many or all of the red posts. Here we instead assume that posts are always hit the average number of times. We are interested in the proportion of red posts active in the 'steady state', in limit of large time.

At time t there are s_t red posts active. Each of these active post emits m sparks, so a total of $s_t \times m$ sparks are emitted. Let each of these sparks be absorbed by a separate post, of which a fraction r/p are red. So at the next timestep, there are $s_{t+1} = s_t mr/p$ red posts active.

The number of active red posts reduces with time if $m < p/r$, and so the arena is extinguished, with $s_\infty = 0$.

The number of active red posts steadily grows with time if $p/r < m$, until there are more sparks emitted than there are posts in total (moving outside our assumption of each spark being absorbed by a separate post), and so the arena saturates with $s_\infty = r$.

The critical value, m_c, where this change of behaviour happens is $m_c = p/r$. Values for m_c for the TFs of interest are shown in figure 8.

Hence the expected behaviour of the single cistrome simulation is to quench for low values of m, saturate for high values of m, and have a tipping point around m_c.

4 Development phase

CoSMoS pattern: *Development:* Build the scientific instrument: produce a simulation platform to perform repeated simulation, based on the output of the *Discovery* phase.

The components of the development phase are:
- *revisit* the Research Context
- develop a *Platform Model*

– develop a *Simulation Platform*
– Argue Instrument Built Appropriately (omitted here)

4.1 Development > revisit

The research context is unchanged in the light of *Discovery* phase activities. The TF concepts need to be reinterpreted in terms of the sparking posts model.

4.2 Development > Platform Modelling

CoSMoS pattern: *Platform Modelling:* From the *Domain Model*, develop a platform model suitable to form the requirements specification for the *Simulation Platform*.

The relevant components of *Platform modelling* are:
– choose a *Modelling Approach* for the platform modelling
– develop the *Platform Model* from the Domain Model
– document *Assumptions* relevant to the platform model

Development > Platform Modelling > Modelling Approach

We use the same approach as for domain model, assisting seamless development.

Development > Platform Modelling > Platform Model

The emergent tipping point behaviour is not part of the platform model. The rest of the 'sparking posts' model carries over from the domain model unchanged.

Instrumentation is added, to collect statistics from the simulator, including post sparking activity. A user interface and visualisation component is added, to control the simulator runs (set the simulation parameters), and examine the output.

Development > Platform Modelling > Assumptions

A.12 the sparks due to an activated post last for one simulation time step
reason simplicity
justification first iteration
consequence half lives and decay rates are not modelled; they may be added in later iterations

4.3 Development > Simulation Platform

CoSMoS pattern: *Simulation Platform:* Develop the executable simulation platform that can be used to run the *Simulation Experiment*.

The relevant components of developing the simulation platform are:
– choose an Implementation Approach
– code and test (details omitted here)

– perform calibration (details omitted here)
– document *Assumptions* relevant to the simulation platform

Development > Simulation Platform > implementation approach

The simulation is implemented as an object-oriented Java application using the MASON simulation environment to handle such things as time-stepping the simulation and on screen graphics (when running in graphical mode).

5 Exploration phase

CoSMoS pattern: *Exploration:* Use the simulation platform resulting from *Development* to explore the scientific questions established during *Discovery*.

The components are:
– *revisit* the Research Context
– perform *Results Modelling*
– perform a *Simulation Experiment*
– Argue Instrument Used Appropriately (omitted here)

5.1 Exploration > revisit

The research context is unchanged in the light of *Discovery* and *Development* phase activities.

5.2 Exploration > Results Modelling

CoSMoS pattern: *Results Modelling:* Develop a results model suitable for interpreting simulation experiment data in Domain Model terms.

The relevant components of results modelling are:
– build a *Visualisation Model*
– build a *Results Model*
– Argue Results Model Appropriate and Consistent (omitted here)

Exploration > Results Modelling > Visualisation Model

CoSMoS pattern: *Visualisation Model:* Visualise the simulation experiment results of the *Data Dictionary* in a manner relevant to the users.

The visualisation mimics the cistrome data in figure 2.

Exploration > Results Modelling > Results Model

The results model is the cistrome activity (number of activated posts) as a function of time.

5.3 Exploration > Simulation Experiment

CoSMoS pattern: *Simulation Experiment:* Use the simulation as a scientific instrument to explore the behaviour of the system.

The relevant components of a simulation experiment are:

- design the experiment
- perform the experiment
- analyse the results

Exploration > Simulation Experiment > design

The parameters p (number of posts) and r (number of red posts) are effectively fixed for any given set of experimentally derived cistrome data (figure 8). We can also generate synthetic data to create systems with a range of p and r values to explore general behaviours.

We identify 4 experiments to perform on the single-arena simulation:

experiment 0 : Effect of m. With p and r fixed and $s_0 = r$, explore the effect of m by locating those values of m for which the system remains fully saturated: all red posts are activated at all time steps. Compare this with the expected m_c value (figure 8) for a noiseless system.

experiment 1 : Effect of s_0, sensitivity to initial conditions. Repeat experiment 0 with smaller values of s_0.

experiment 2 : Effect of r. Create arenas with a fixed p and a range of r values. At each value of r, determine the values of m for which the system remains saturated throughout the simulation.

experiment 3 : Effect of noise. Keeping the ratio of p to r fixed at the value in the biological data, investigate the effect of reducing p. This will give some insight into how the data scales up within the context of our model, and whether we can use smaller arenas in experiments to improve simulation performance.

Number of simulation runs. We are not performing any statistical analyses at this stage of the project, merely inspecting behaviour. However, the simulation is essentially stochastic, and when we do come to perform statistics, we will need to choose the number of runs based on the significance, power, and effect size of interest. For consistency, we make that choice now, and use the relevant number of runs.

We require a statistical significance of 99% (a 1% false positive rate), a statistical power of 99% (a 1% false negative rate), and a 'medium' effect size (Cohen's

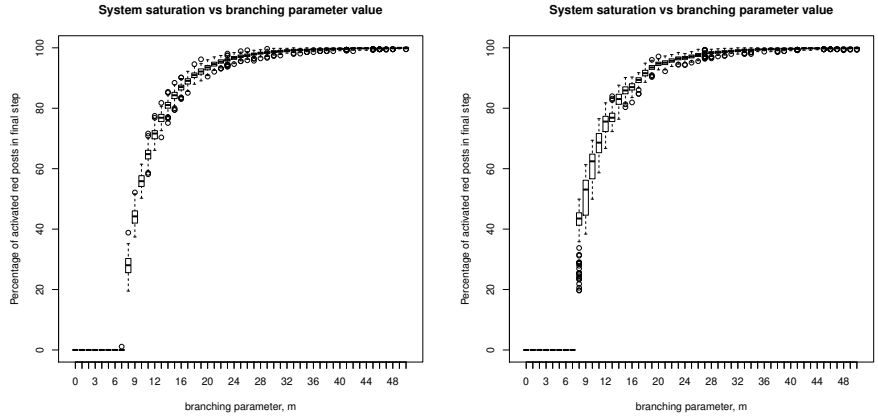

Fig. 9. p and r corresponding to Nanog data; (left) experiment 0: $s_0 = r$; (right) experiment 1: $s_0 = r/2$. Recall $m_c = 6.8$

$d = 0.5$, the ability to distinguish a difference in means of 0.5 of a standard deviation). Calculating the required sample size for these experimental parameters[3] gives 192.

We round this up, and take the number of runs to be $N = 200$.

Protocol. One simulation run comprises the p and r values of a particular arena (chosen to match Nanog, Sox2, Oct4 data), an m value (1–50), and a starting activity ($s_0 = r$ for experiment 0; $s_0 = r/2$ for experiment 1).

For each simulation run, we record the proportion of active red posts at the final timestep, $T = 1000$.

For each parameter set (p, r, m, s_0), we run the simulation $N = 200$ times.

5.4 Exploration > Simulation Experiment > analyse results

Experiments 0 and 1

Experiment 0 uses $s_0 = r$: all red posts initially active. Experiment 1 uses $s_0 = r/2$: half the red posts initially active.

See figures 9–11 for the results of the simulation runs.

The observed values of m where the system 'switches on', and can maintain saturation, are close to the calculated m_c values. However, m has to be somewhat higher than this to saturate the finite-sized arena.

Starting with only half the posts active makes little difference to the results.

[3] using, for example, the calculator at http://powerandsamplesize.com/Calculators/Compare-2-Means/2-Sample-Equality

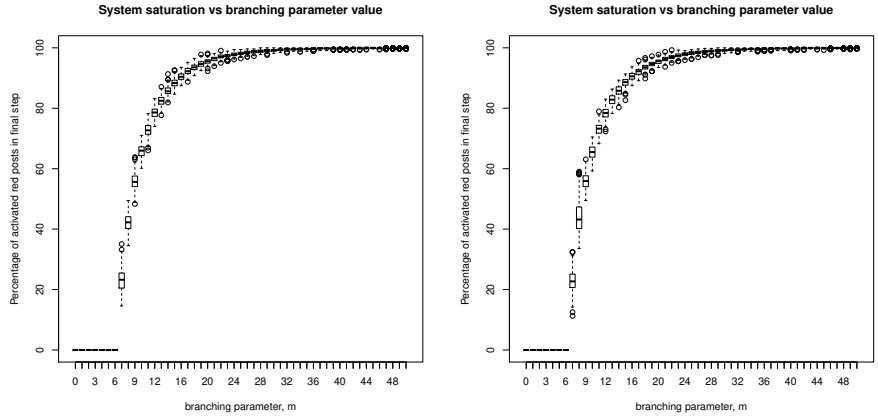

Fig. 10. p and r corresponding to Sox2 data; (left) experiment 0: $s_0 = r$; (right) experiment 1: $s_0 = r/2$. Recall $m_c = 6.1$

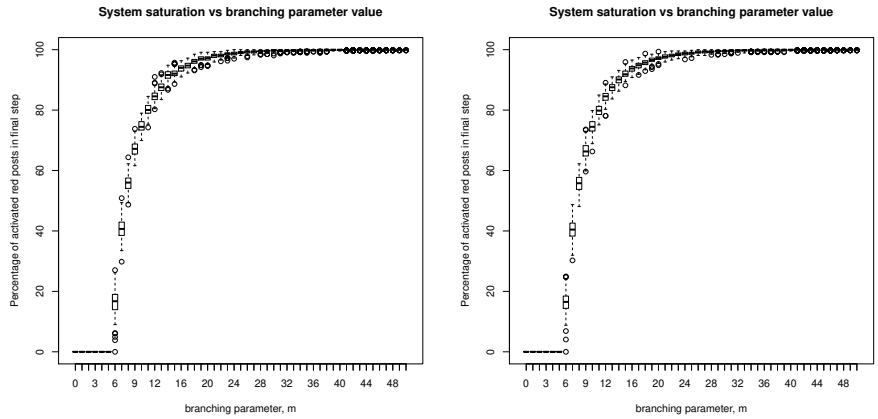

Fig. 11. p and r corresponding to Oct4 data; (left) experiment 0: $s_0 = r$; (right) experiment 1: $s_0 = r/2$. Recall $m_c = 5.5$

Experiment 2

For experiment 2, we took $p = 4310$ (as in Nanog), and $r = 200, 400, 600, 800$, to see how the value of m_c changes. We used $s_0 = r$ throughout.

See figures 12–13 for the results of the simulation runs.

Recall that the theoretical tiping point value is $m_c = p/r$. So as r increases, m_c should decrease. This is observed (figure 12).

Also note that the smaller r, the noisier the behaviour. This is expected as stochastic effects will be more prominent when there are fewer red posts available.

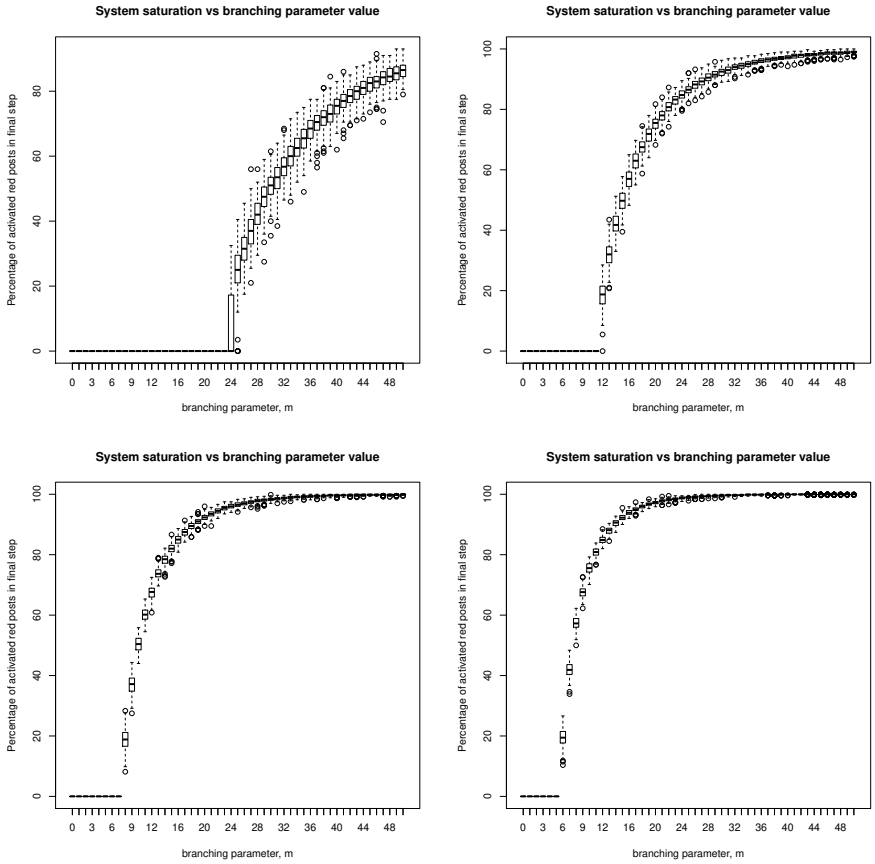

Fig. 12. Experiment 2: varying r; here $p = 4310$: (top left) $r = 200$; (top right) $r = 400$; (bottom left) $r = 600$; (bottom right) $r = 800$

r	m obs	m_c
200	23–24	21.6
400	11-12	10.8
600	7–8	7.2
800	5–6	5.4

Fig. 13. Experient 2: observed value of m at tipping point, versus calculated value m_c

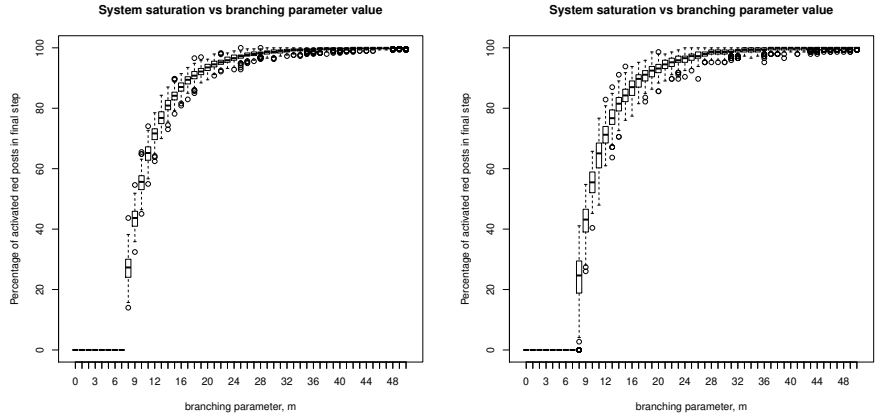

Fig. 14. Experiment 3: varying p with constant p/r: (left) $p = 2000, r = 293$; (right) $p = 1000, r = 146$

Experiment 3

For experiment 3, we took $p/r = 4310/631$ (as in Nanog), and reduced p keeping p/r constant (mimicking a smaller arena but with the same density of red posts). We used $s_0 = r$ throughout.

See figure 14 for the results of the simulation runs; compare with figure 9(top) for the 'full' arena.

The systems tip at the same point, but the behaviour gets noisier as p (and hence r) decreases, and stochastic effects become more pronounced.

6 Discussion

This paper documents and illustrates the use of CoSMoS patterns to perform a complete iteration of a CoSMoS simulation project, from initial discovery, through development, to exploration. There were several lessons learned, summarised here.

It is not always clear whether information should be included in the Domain, or Domain Model, sections, particularly relating to assumptions. However, it is more important to document the information that to agonise over precisely which section to document it in.

Not all patterns are applicable. For example, here the Domain Model Cartoon had to be presented within the Domain Model section, rather than as a prior illustration. Additionally, the TF BP model is so abstracted from the Domain, that aspects such as the Domain Experiment Model [Andrews and Stepney, 2015] are not relevant, and so have been omitted. Again, it is more important to follow the spirit of the CoSMoS approach rather than the letter of every pattern.

Not every aspect of the CoSMoS approach needs to be performed with complete rigour. This simulation is not safety critical, so some aspects have been omitted (such as justification of all assumptions, and argumentation of fitness-for-purpose). The extra effort needed to complete all aspects should be expended only if it gives benefit.

Although the presentation is sequential and hierarchical, the historical process was not. We spent many short iterations, and considerable backtracking (for example, see figure 3), before finally fixing on the 'sparking posts' model. The CoSMoS patterns define what information should be recorded by the end of the project, but not the order it needs to be produced. Some uses of CoSMoS can apply the patterns in significantly different orders, for example [Andrews and Stepney, 2014].

We might not have arrived at the conceptual sparking posts model without taking an iterative approach. The need to have just a single-cistrome model for this first iteration revealed a fundamental misunderstanding that the modellers were having about the background TF BP model.

Although we were taking an agile approach, producing minimal simulation models and code, collaborations meetings would often generate interesting but out of current scope ideas. We invented the concept of the "to don't" list: a place to record the ideas for future reference, in a manner that made it clear they were not to be included in the current iteration. Some of these ideas also prompted the recognition of assumptions in the current iteration.

The Domain Scientist (Halley) was new to the CoSMoS approach at the start of the project, but had previous experience working with modellers using different approaches on other projects. Halley reports that CoSMoS is a flexible tool to produce objective scientific simulations, and allows progress without being funnelled into preconceptions imposed by a specific toolset or implementation approach.

7 Summary, Conclusions, Future Work

This work has run through a complete CoSMoS cycle, producing the first iteration of the system: a single cistrome model.

The results demonstrate that the single-cistrome model exhibits its tipping point close to the predicted value of m_c, but the tipping is not particularly sharp, so for values of m close to m_c, there is a lot of noise in the system.

In order to generate results that have genuine biological relevance, it will be necessary to create a simulation of two or more cistromes interacting with each other via the TFs that each produces. For example, we will investigate model behaviour when the Oct4, Sox2 and Nanog branching processes are allowed to interact. Given the groundwork developed in this first iteration, the modelling and simulation work in for the second iteration, to augment the system with multiple cistromes, should be relatively straightforward. We are currently developing this second iteration.

Beyond this, future iterations could include:

– More complex connections within networks of cistromes, including inhibition and negative feedback, combinatorial binding of TFs, and indicators of 3D genomic or chromosomal architecture. The inclusion of inhibition of gene expression is particularly relevant to the process of pluripotency exit, as batteries of differentiation genes are suddenly expressed.
– A Domain Specific Language with which we can describe the network
– TF half life variability
– Epigenetic histone marks that may help to shape circuitry self-organisation
– Combinatorial binding of TFs to enhancer sites that impart transcriptional synergy [Struhl, 2001]
– Multicellular model incorporating cell-cell signalling

The model presented here represents a novel example of self-organisation that may apply to other complex systems. It is of interest from a purely theoretical perspective because it helps to demonstrate how distributed interactions among units result in higher ordered emergent behaviours. Such complexity could provide dynamic templates of organisation upon which natural selection builds additional elaborations [Halley and Winkler, 2008].

Acknowledgments

This work was performed as part of the CellBranch project, funded by the UK's Biotechnology and Biological Sciences Research Council (BBSRC), project reference BB/L018705/1.

References

Alexander, C. et al. (1977). *A Pattern Language: towns, buildings, construction.* Oxford University Press.

Anderson, P. W. (1991). Is complexity physics? is it science? what is it? *Physics Today*, 44(7):9.

Andrews, P. S., Polack, F. A. C., Sampson, A. T., Stepney, S., and Timmis, J. (2010). The CoSMoS process, version 0.1: A process for the modelling and simulation of complex systems. Technical Report YCS-2010-453, Department of Computer Science, University of York.

Andrews, P. S. and Stepney, S. (2014). Using CoSMoS to reverse engineer a domain model for Aevol. In *Proceedings of the 2014 Workshop on Complex Systems Modelling and Simulation, New York, USA, July 2014*, pages 61–79. Luniver Press.

Andrews, P. S. and Stepney, S. (2015). The CoSMoS Domain Experiment Model. In *Proceedings of the 2015 Workshop on Complex Systems Modelling and Simulation, York, UK, July 2015*. Luniver Press.

Andrews, P. S., Stepney, S., and Timmis, J. (2012). Simulation as a scientific instrument. In Stepney et al. [2012], pages 1–10.

Ay, A. and Arnosti, D. N. (2011). Mathematical modeling of gene expression: a guide for the perplexed biologist. *Critical Reviews in Biochemistry and Molecular Biology*, 46(2):137–151.

Bak, P. and Paczuski, M. (1993). Why nature is complex. *Physics World*, 6(12):39–43.

Bak, P. and Paczuski, M. (1995). Complexity, contingency, and criticality. *Proceedings of the National Academy of Science USA*, 92:6689–6696.

Ball, P. (1999). Transitions still to be made. *Nature*, 402:C73–C76.

Ball, P. (2001). *The Self-Made Tapestry*. Oxford University Press.

Bornholdt, S. (2005). Less is more in modeling large genetic networks. *Science*, 310:449–451.

Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., Guenther, M. G., Kumar, R. M., Murray, H. L., Jenner, R. G., Gifford, D. K., Melton, D. A., Jaenisch, R., and Young, R. A. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, 122:947–956.

Bray, D. (2003). Molecular networks: the top-down view. *Science*, 301:1864–1865.

Chaisson, E. (2004). Complexity: An energetics agenda. *Complexity*, 9(3):14–21.

Crutchfield, J. P., Farmer, J. D., Packard, N. H., and Shaw, R. S. (1986). Chaos. *Scientific American*, 255(6):38–49.

Driscoll, M. E. (2009). Is big data at a tipping point? http://www.analyticbridge.com/profiles/blogs/is-big-data-at-a-tipping-point. [Accessed: 2015-05-01].

Ekeland, I. (2002). In the balance. *Nature*, 417:385.

Farmer, J. D. and Packard, N. H. (1986). Evolution, games, and learning: Models for adaptation in machines and nature. An introduction to the proceedings of the CNLS Conference, Los Alamos, May 1985. *Physica D*, 22:vii–xii.

Ferrell, J. (2009). Q&A: Systems biology. *Journal of Biology*, 28.

Gallagher, R. and Appenzeller, T. (1999). Beyond reductionism. *Science*, 284(5411):79.

Gollub, J. P. and Langer, J. S. (1999). Pattern formation in nonequilibrium physics. *Reviews of Modern Physics*, 71(2):S396–S403.

Halley, J. D., Burden, F. R., and Winkler, D. A. (2009). Stem cell decision making and critical-like exploratory networks. *Stem Cell Research*, 2(3):165–177.

Halley, J. D., Smith-Miles, K., et al. (2012). Self-organizing circuitry and emergent computation in mouse embryonic stem cells. *Stem Cell Research*, 8(2):324–333.

Halley, J. D. and Winkler, D. A. (2008). Critical-like self-organization and natural selection: Two facets of a single evolutionary process? *BioSystems*, 92(2):148–158.

Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular to modular cell biology. *Nature*, 402:C47–C52.

Howe, D., Costanzo, M., et al. (2008). Big data: The future of biocuration. *Nature*, 455(7209):47–50.

Kadanoff, L. P. (1987). Chaos: A view of complexity in the physical sciences. In *From Order to Chaos II Essays: Critical Chaotic and Otherwise*. World Scientific.

Kalkan, T. and Smith, A. (2014). Mapping the route from naive pluripotency to lineage specification. *Phil. Trans. R. Soc. B*, 369:20130540.

Kauffman, S. (1995). *At Home in the Universe*. Oxford University Press.

Kim, J., Chu, J., Shen, X., Wang, J., and Orkin, S. H. (2008). An extended transcriptional network for pluripotency of embryonic stem cells. *Cell*, 132(6):1049–1061.

Lander, A. D. (2010). The edges of understanding. *BMC Biology*, 8(1):40.

Loh, Y.-H., Wu, Q., Chew, J.-L., Vega, V. B., Zhang, W., Chen, X., Bourque, G., George, J., Leong, B., Liu, J., Wong, K.-Y., Sung, K. W., Lee, C. W. H., Zhao, X.-D., Chiu, K.-P., Lipovich, L., Kuznetsov, V. A., Robson, P., Stanton, L. W., Wei, C.-L., Ruan, Y., Lim, B., and Ng, H.-H. (2006). The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nature Genetics*, 38(4):431–440.

MacArthur, B. D., Maayan, A., and Lemischka, I. R. (2008). Toward stem cell systems biology: From molecules to networks and landscapes. *Cold Spring Harbor Symposia on Quantitative Biology*, 73:211–215.

Martello, G. and Smith, A. (2014). The nature of embryonic stem cells. *Annual Review of Cell and Developmental Biology*, 30:647–675.

Mesarovic, M. D., Sreenath, S. N., and Keene, J. D. (2004). Search for organising principles: understanding in systems biology. *Systems Biology*, 1(1):19–27.

Muñoz Descalzo, S., Rué, P., et al. (2013). A competitive protein interaction network buffers Oct4-mediated differentiation to promote pluripotency in embryonic stem cells. *Molecular Systems Biology*, 9:694.

Nichols, J. and Smith, A. (2009). Naive and primed pluripotent states. *Cell Stem Cell*, 4(6):487–492.

Nicolis, G. and Prigogine, I. (1977). *Self-Organization in Nonequilibrium Systems*. John Wiley & Sons.

Niwa, H., Miyazaki, J.-i., and Smith, A. G. (2000). Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. *Nature Genetics*, 24(4):372–6.

Parisi, G. (1993). Statistical physics and biology. *Physics World*, 6:42–47.

Stepney, S. (2012). A pattern language for scientific simulations. In Stepney et al. [2012], pages 77–103.

Stepney, S. and Andrews, P. S. (2015). CoSMoS special issue editorial. *Natural Computing*, 14:1–6.

Stepney, S., Andrews, P. S., and Read, M., editors (2012). *Proceedings of the 2012 Workshop on Complex Systems Modelling and Simulation, Orleans, France, September 2012*. Luniver Press.

Stepney, S. et al. (2016). *Engineering Simulations as Scientific Instruments*. Springer. [in prep].

Struhl, K. (2001). Gene regulation. a paradigm for precision. *Science*, 293:10541055.

Teles, J., Pina, C., et al. (2013). Transcriptional regulation of lineage commitment – a stochastic model of cell fate decisions. *PLOS Computational Biology*, 9(8):e1003197.

Vicsek, T. (2002). The bigger picture. *Nature*, page 131.

Weatherall, D. J. (2001). Phenotype-genotype relationship in monogenic disease: lessons from the Thalassemias. *Nature Reviews Genetics*, 2:245–255.

Xu, H., Schaniel, C., Lemischka, I. R., and Ma'ayan, A. (2010). Toward a complete in silico, multi-layered embryonic stem cell regulatory network. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 2(6):708–733.

Zandstra, P. and Clarke, G. (2014). Computational modeling and stem cell engineering. In Nerem, R. M., Loring, J., et al., editors, *Stem Cell Engineering*, pages 65–97. Springer.