Waddle:
Waddington Epigenetic Landscapes

Felicia Burtscher, Lucas Ducrot, Madeleine Hall, Luis Torada

MSc in Bioinformatics and Theoretical Systems Biology
Imperial College London

17th April 2018

Outline

1. Motivation, Introduction & Workflow [FB]
2. Model input & Action-Based Method (ABM) [LT]
3. Simulations, Probability Flux Method & Kernel Density Estimation, 2D-Model Results [LD]
4. Data, Discussion & Conclusion [MH]

Motivation, Introduction & Workflow [FB]

Motivation

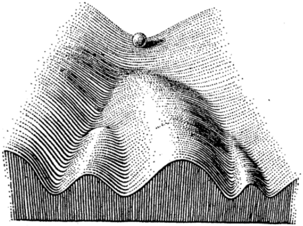


Figure 1: *Part of an Epigenetic Landscape* (Waddington, 1940)

"The path followed by the ball [. . .] corresponds to the developmental history of a particular part of the egg. There is first an alternative, towards the right or the left. Along the former path, a second alternative is offered; along the path to the left, the main channel continues leftwards, but there is an alternative path which, however, can only be reached over a threshold" (Waddington, 1940).

Introduction: Stochastic models

From a stochastic system to an epigenetic landscape:

- Some stochastic systems can be expressed in terms of a potential.
- Potential functions are similar to epigenetic landscapes.

A general stochastic model is of the form:

$$dX_t = f(X_t, t)dt + g(X_t, t)dW_t \tag{1}$$

and can be written as an ODE

$$\frac{dX_t}{dt} = f(X_t, t)$$

in the limit of low noise (high copy numbers).

Introduction: Landscape formation

Decomposition of the forcing vector $f(X;t)$:

$$f(X;t) = -\nabla U(X;t) + f_U(X;t) \tag{2}$$

where $\nabla U(X;t)$ is the gradient of a potential, and $f_U(X;t)$ is the remaining component, often referred to as the *curl*.

We call U the *quasi-potential* and it is analogous to the epigenetic landscape. f_U is the remainder term, and fills out remaining dynamics.

Introduction: Define our goal

- Literature review: Models like *NetLand* (in Java) already available.
- Focus on visualisation by last year's group.

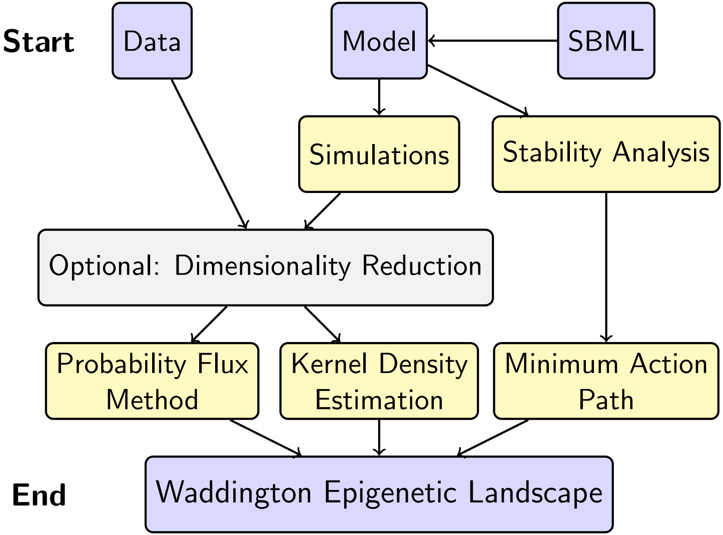
Our focus:

- Methodology, less on visualisation
- in Julia

Goal

Provide different tools and methods in Julia to (1) simulate and analyse the landscape and (2) visualise certain genes or gene combinations (after applying dimensionality reduction) given a single-cell data set.

Workflow



Start

Data

Model

SBML

Simulations

Stability Analysis

Optional: Dimensionality Reduction

Probability Flux Method

Kernel Density Estimation

Minimum Action Path

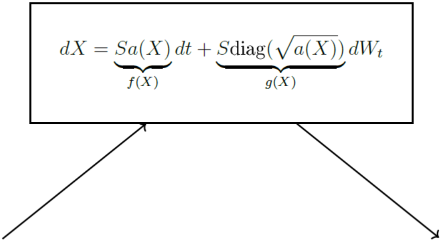
End

Waddington Epigenetic Landscape

Model input & Action-Based Method (ABM) [LT]

Stochastic Differential Equations (SDE) model

Stochastic Differential Equations (SDE) model:



$$\boxed{dX = \underbrace{Sa(X)}_{f(X)}\,dt + \underbrace{S\mathrm{diag}(\sqrt{a(X)})}_{g(X)}\,dW_t}$$

Possible sources: manual, SBML file

Utility: simulations (PFM), Action Based Method (ABM)

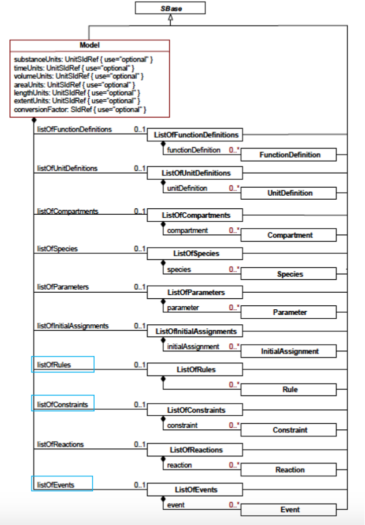
Possible sources: SBML file

1. **General structure:**

   XML-based
   Model object that inherits lists.

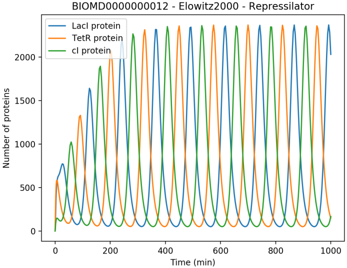
2. **Approach:**

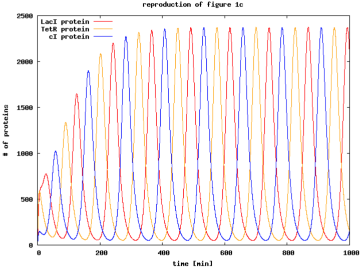
   Strings – metaprogramming → Julia ODEs.
   Limited complexity (exceptions identifiable).

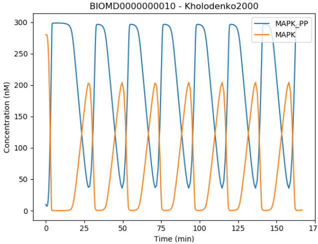


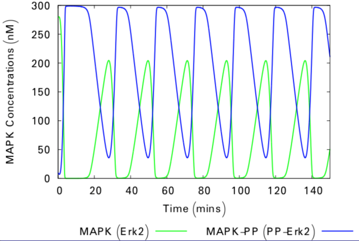
Possible sources: SBML file

3. **Validation:**


BIOMD0000000012 - Elowitz2000 - Repressilator


reproduction of figure 1c


BIOMD0000000010 - Kholodenko2000



Utility: Action-Based Method (ABM)

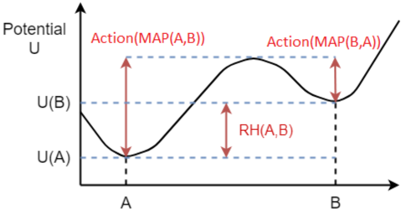
Large deviations Theory:

Action function:

$$S(\varphi, T) = \int_0^T \sum_i (\dot{x}_i - f_i(x))^2 / g_i^2(x) dt \tag{3}$$

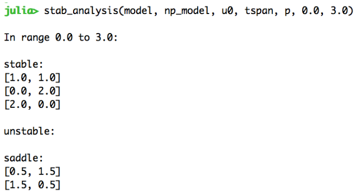
Quasi-potential barrier, S(minimum-action path):

$$U(x_1, x_2) = \inf_{T>0} \inf_{\varphi \in \bar{C}_{x_1}^{x_2}(0,T)} S_T(\varphi) \tag{4}$$

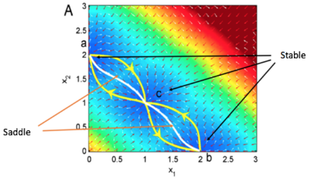


Utility: Action-Based Method (ABM)

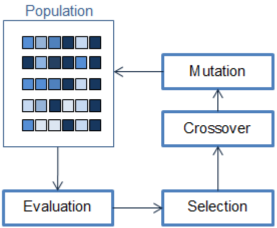
Stability Analysis:



```
julia> stab_analysis(model, np_model, u0, tspan, p, 0.0, 3.0)

In range 0.0 to 3.0:

stable:
[1.0, 1.0]
[0.0, 2.0]
[2.0, 0.0]

unstable:

saddle:
[0.5, 1.5]
[1.5, 0.5]
```



Genetic Algorithm:



Population

Mutation

Crossover

Evaluation

Selection

Simulations, Probability Flux Method & Kernel Density Estimation, 2D-Model Results [LD]

Landscapes from Simulations of a SDE model

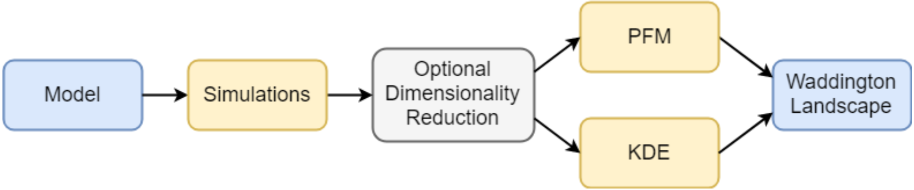


Figure 2: Workflow using simulations to build lanscapes

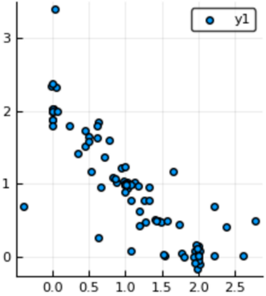
Simulations

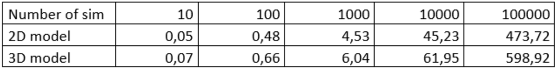
The analyzed model is a SDE:

$$dX = f(X)dt + g(X)dW_t$$

From this model, using the DifferentialEquations package of Julia, the method runs simulations:

- random initial points
- fix amount of time





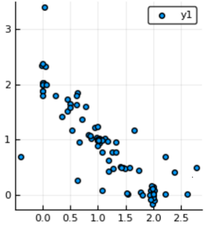
Number of sim	10	100	1000	10000	100000
2D model	0,05	0,48	4,53	45,23	473,72
3D model	0,07	0,66	6,04	61,95	598,92

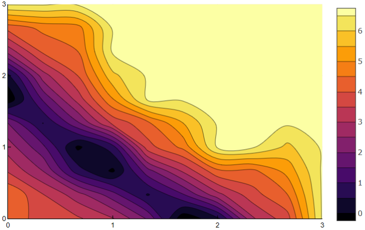
Probability Flux Method

The idea of the PFM is to compute the potential landscape U thanks to:

$$U \propto \ln(P_s)$$

where P_s is the density function calculated from the simulations
We discretize the space and calculate the density distribution.





Main problems: Requires lot of simulations, does not describe well the low probability area of the space.

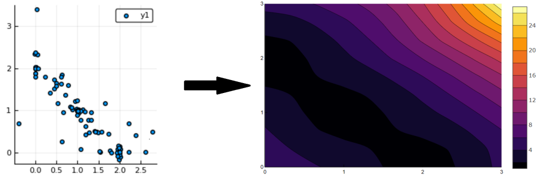
Kernel Density Estimation

KDE is a variant of the PFM where P_s is calculated as a parameterized function $\hat{f}_h$ where x_i are the results of the simulations and h a bandwidth:

$$\hat{f}_h(x) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right)$$

Here K is a kernel function, the Gaussian Kernel is the most widely used:

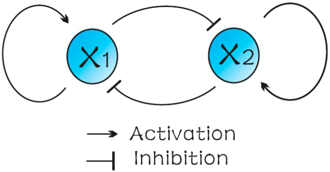
$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$



Main issue is the choice of the optimal bandwidth.

Test on a 2D model

The 2-dimensional model we use to test our methods is from a basic gene regulatory network of 2 nodes. These genes are self activating and mutually inhibiting each other.



$$dX_1 = \left(\frac{a_1 X_1^n}{S^n + X_1^n} + \frac{b_1 S^n}{S^n + X_2^n} - k_1 X_1 \right) dt + \sqrt{\left| \frac{a_1 X_1^n}{S^n + X_1^n} + \frac{b_1 S^n}{S^n + X_2^n} - k_1 X_1 \right|} dW_t$$

$$dX_2 = \left(\frac{a_2 X_2^n}{S^n + x_2^n} + \frac{b_2 S^n}{S^n + X_1^n} - k_2 X_2 \right) dt + \sqrt{\left| \frac{a_2 X_2^n}{S^n + x_2^n} + \frac{b_2 S^n}{S^n + X_1^n} - k_2 X_2 \right|} dW_t$$

Results of the 2D model

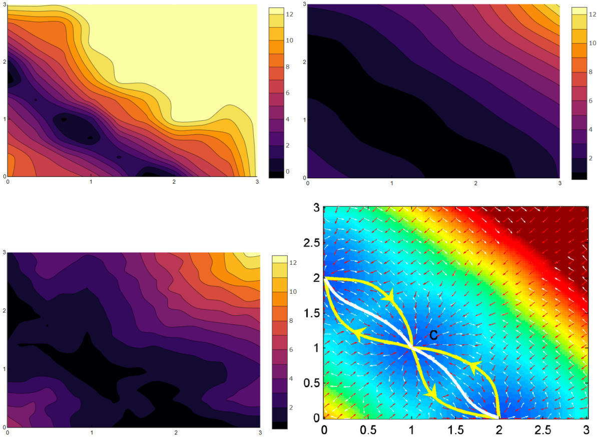


Figure 3: Colored maps of the 2D model potential computed from PFM (top-left), PFM-KDE (top-right), ABM (bottom-left) and control plot (bottom-right). X1 in X-axis and X2 in Y-axis.

Data, Discussion & Conclusion [MH]

Landscapes from Data

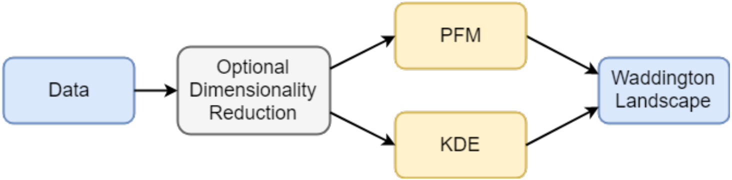


Figure 4: Workflow using data to build landscapes

- Data points can be thought of as equivalent to end states from simulations of models
- We apply the same approach to generate landscapes

Landscapes from Data

- 547 cells, 96 genes[1]: seperated by time (0, 24, . . . , 168 hours) or type (ESC, EPI, NPC)
- Dimensionality Reduction:
  - Feature selection: highest correlation, information theory measures
  - Feature extraction: PCA, PPCA, etc.

    Loss of biological relevance

⇒ 2 dimensions to generate landscape

[1] Neil Smyth, University of Southampton

Landscapes from Data

- Feature selection/extraction can determine the most informative landscapes
- Common form of measurement noise in this type of data is false zeros

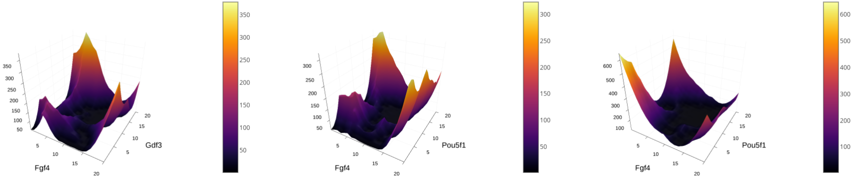


Figure 5: Landscapes - highest correlation, highest MI, zeros removed.

Landscapes from Data

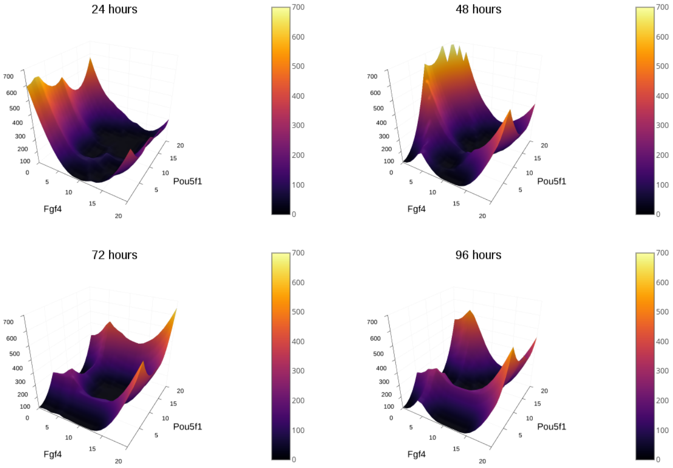


Figure 6: Separation of data by time allows visualisation of landscape evolution. Animations can be generated to show moving landscapes.

Discussion

- Julia: fast, clear mathematical syntax, growing availability of libraries (key to this project - DifferentialEquations, MultivariateStats, Plots), but still developing
- Potential to develop techniques further to incorporate higher dimensional systems and datasets

Summary

- Developed tools to construct Waddington landscapes in Julia
- Inputs: SBML file, model, data
- Methods: ABM, PFM, KDE
- Provide means for further exploration of landscapes, including stability analysis and dimensionality reduction