Stochastic modelling of genetic interaction in budding yeast

Darren Wilkinson
http://tinyurl.com/darrenjw
School of Mathematics & Statistics,
Newcastle University, UK

2013 S&N Meeting
Newcastle University
24th May, 2013
Overview

- Background: Budding yeast as a model for genetics
- High-throughput robotic genetic experiments
- Image analysis and data processing
- Stochastic modelling of growth curves
- Hierarchical modelling of genetic interaction
- Summary and conclusions

Joint work with Jonathan Heydari, Conor Lawless and David Lydall (and others in the “Lydall lab”)
Saccharomyces cerevisiae

- *Saccharomyces cerevisiae*, often known as budding yeast, and sometimes as brewer’s yeast or baker’s yeast, is a single-celled eukaryotic organism.
- Eukaryotic cells contain a nucleus (and typically other organelles, such as mitochondria).
- Budding yeast is an interesting species commercially, due to its use in baking, beer brewing, wine making, yeast extract, etc.
- It is also useful as a model for higher eukaryotes, having a great deal of biological function conserved with humans.
- It is the most heavily studied and well-characterised model organism in biology (e.g. first fully sequenced eukaryote).
Budding yeast under the microscope
Possible to obtain a library of around 4,500 mutant strains, each of which has one of the non-essential genes silenced through insertion of a (kanMX) antibiotic resistance cassette and tagged with a unique DNA barcode.

These strains (stored frozen in 96-well plates) can be manipulated by robots in 96-well plates (8×12), or on solid agar in 96, 384 or 1536-spot format.

**Synthetic Genetic Array (SGA)** is a clever genetic procedure using robots to systematically introduce an additional mutation into each strain in the library by starting from a specially constructed query strain containing the new mutation.
Telomeres

- The ends of linear chromosomes require special protection in order not to be targeted by DNA damage repair machinery (bacteria often avoid this problem by having just one chromosome arranged in a single loop)
- Telomeres are the ends of the chromosomal DNA (which have a special sequence), bound with special telomere-capping proteins that protect the telomeres
- CDC13 is an essential telomere-capping protein in yeast
- *cdc13-1* is a point-mutation of *cdc13*, encoding a temperature-sensitive protein which functions similarly to wild-type CDC13 below around 25 degrees Celsius, and leads to “telomere-uncapping” above this temperature
David Lydall’s (budding) yeast lab is interested in using a range of HTP technologies for genome-wide screening for interactions relevant to DNA damage response and repair pathways, with a particular emphasis on telomere maintenance. Much of this work centres around the use of robotic protocols in conjunction with genome-wide knockout libraries and synthetic genetic array (SGA) technology to screen for genetic interactions with known telomere maintenance genes.
Basic structure of an experiment

1. Introduce a **mutation** (such as *cdc13-1*) into an SGA query strain, and then use SGA technology (and a robot) to **cross** this strain with the single deletion library in order to obtain a new library of double mutants.

2. **Inoculate** the strains into liquid media, grow up to saturation then spot back on to solid agar 4 times.

3. **Incubate** the 4 different copies at different temperatures (treatments), and image the plates multiple times to see how quickly the different strains are growing.

4. Repeat steps 2 and 3 four times (to get some idea of experimental variation).

5. Repeat steps 2 to 4 with a “control” library that does not include the query mutation.
Some numbers relating to an experiment

- Initial SGA work (introducing mutations into the query and the library) takes around 1 month of calendar time, and several days of robot time.
- The inoculation, spotting and imaging of the 8 repeats takes 1 month of calendar time, and around 2 weeks of robot time.
- The experiment uses around £5,000 of consumables (plastics and media).
- The library is distributed across 60 96-well plates or 15 solid agar plates (in 384 format, or 1536 in quadruplicate).
- If each plate is imaged 40 times, there will be around 40k high-resolution photographs of plates in 384 format, corresponding to around 15 million colony growth measurements (400k time series).
- This is big data!
The colony handling robot
Data analysis pipeline

- **Image processing** (from images to colony size measurements)
- **Fitness modelling** (from colony size growth curves to strain fitness measures)
- **Modelling genetic interaction** (from strain fitness measures to identification of genetically interacting strains, ranked by effect size)

Possible to carry out three stages separately, but benefits to joint modelling through borrowed strength and proper propagation of uncertainty. Not practical to integrate image processing step into the joint model, but possible to jointly model second two stages.
Automated image analysis (Colonyzer)
Growth curve

A

B

Growth curve modelling

- We want something between a simple smoothing of the data and a detailed model of yeast cell growth and division.
- **Logistic growth models** are ideal — simple semi-mechanistic models with interpretable parameters related to strain fitness.
- Basic deterministic model:
  \[ \frac{dx}{dt} = rx(1 - x/K), \]
  subject to initial condition \( x = P \) at \( t = 0 \)
- \( r \) is the **growth rate** and \( K \) is the **carrying capacity**
- Analytic solution:
  \[ x(t) = \frac{KP e^{rt}}{K + P(e^{rt} - 1)} \]
Statistical model

- Model observational measurements \( \{Y_{t1}, Y_{t2}, \ldots \} \) with

\[
Y_{ti} = x_{ti} + \varepsilon_{ti}
\]

- Can fit to observed data \( y_{ti} \) using non-linear least squares or MCMC

- Can fit all (400k) time courses simultaneously in a large hierarchical model which effectively borrows strength, especially across repeats, but also across genes

- Generally works well (fine for most of the downstream scientific applications), but fit is often far from perfect...
Fitting the logistic curve

![Graph showing colony size over time for YAL003W yeast.](image-url)
Improved modelling of colony growth curves

- Could use a **generalised logistic model** (Richards’ curve) which breaks the symmetry in the shape of “take off” and “landing”

\[ \frac{dx}{dt} = rx \left( 1 - \left( \frac{x}{K} \right)^\nu \right) \]

- This helps, but doesn’t address the real problem of strongly auto-correlated residuals
- Better to introduce noise into the dynamics to get a logistic growth diffusion process
Stochastic logistic growth diffusion

- Well-known stochastic generalisation of the logistic growth equation, expressed as an Itô stochastic differential equation (SDE):

\[ dX_t = rX_t(1 - X_t/K)dt + \xi^{-1/2}X_t \, dW_t \]

- The **drift** is exactly as for the deterministic model
- The **diffusion** term injects some noise into the dynamics
- The **multiplicative noise** ensures that this defines a non-negative stochastic process
Sample trajectories from the logistic diffusion

Stochastic logistic growth

Darren Wilkinson — S&N, 24/5/2013

Stochastic modelling of genetic interaction in budding yeast
Statistical model

- Model observational measurements $\{Y_{t_1}, Y_{t_2}, \ldots\}$ with

\[ Y_{t_i} = X_{t_i} + \varepsilon_{t_i} \]

where $X_{t_i}$ refers to our realisation of the diffusion process

- Need somewhat sophisticated algorithms to fit these sorts of SDE models to discrete time data

- Standard algorithms would require knowledge of the transition kernel of the diffusion process, but this is not available for the logistic diffusion

- Lots of work on Bayesian inference for intractable diffusions (Golightly & W, ’05, ’06, ’08, ’10, ’11), but this won’t scale to simultaneous fitting of tens of thousands of realisations
Approximating the stochastic logistic diffusion

- Computational constraints mean that we can only really consider working with diffusions having tractable transition kernels (as then we can apply standard MCMC methods for discrete time problems)
- Would therefore like a tractable approximation to the stochastic logistic diffusion
- Román–Román & Torres–Ruiz (2012) propose just such an approximation:

\[ dX_t = \frac{br}{e^{rt} + b} X_t \, dt + \xi^{-1/2} X_t \, dW_t, \]

where \( b = (K/P) - 1 \), and use it to fit measured growth curves to data
The Román–Román & Torres–Ruiz (RRTR) model

- The RRTR model has a tractable transition kernel with log-normal increments, so is convenient for fitting to data.
- Where does it come from?
  - Comparing against the logistic diffusion, we see that the expression \( r(1 - X_t/K) \) has been replaced with \( \frac{br}{e^{rt}+b} \), which is exactly what it would be in the deterministic case.
  - This linearises the SDE, rendering it tractable.
- Unfortunately this approximation isn’t particularly good...
The logistic diffusion and the RRTR approximation

The RRTR model behaves asymptotically like Geometric Brownian motion (GBM), where as the true process is mean reverting.
The RRTR model has the desirable feature of log-normal increments, but has problems with long-term behaviour.

Alternatively, if we apply a log transformation to the logistic diffusion and then carry out a linear noise approximation, the result will also be a process with log-normal increments, but will have mean-reverting behaviour which is clearly desirable here.

Putting $U_t = \log X_t$, Itô’s formula gives

$$dU_t = \left(r - \frac{1}{2}\xi - \frac{r}{K}e^{U_t}\right)dt + \xi^{-1/2}dW_t$$
Linear noise approximation (LNA)

- Decompose $U_t$ into a deterministic component and a stochastic residual process

$$U_t = v_t + Z_t$$

where $v_t$ solves the deterministic part

$$\frac{dv_t}{dt} = r - \frac{1}{2\xi} - \frac{r}{K}e^{v_t}$$

- Subtracting out the deterministic solution from $U_t$ leaves a residual process of the form

$$dZ_t = \frac{r}{K}e^{v_t}(1 - e^{Z_t})dt + \xi^{-1/2}dW_t$$
Linear noise approximation (LNA)

- Applying the linear approximation $1 - e^{Z_t} \approx -Z_t$ to linearise the drift gives

$$dZ_t = -\frac{r}{K} e^{v_t} Z_t dt + \xi^{-1/2} dW_t$$

- Substituting in for $v_t$ then gives

$$dZ_t = -\frac{abPe^{at}}{bP(e^{at} - 1) + a} Z_t dt + \xi^{-1/2} dW_t,$$

where $a = r - 1/(2\xi)$ and $b = r/K$

- This is a (zero) mean-reverting time-varying Ornstein–Uhlenbeck (OU) process, and can be solved exactly, giving a normal transition kernel.
The (log) LNA is a very good approximation to the true process, with tractable log-normal increments.
Further simplifications and approximations

- The LNA is a good model with a tractable transition kernel
- We can implement standard discrete time MCMC methods to estimate model parameters together with the unobserved latent trajectories
- Embedding in a hierarchical model is straightforward
- These methods work fine for hundreds of growth curves, but are still problematic for tens of thousands of growth curves
- Ideally we would like to integrate out the latent process
  - If we are prepared to assume linear Gaussian error on the log scale, we can use Kalman filtering techniques to integrate out the latent process (but this isn’t very plausible)
  - Alternatively, we could apply a LNA directly to the logistic diffusion (without first transforming), and assume linear Gaussian error on that scale (ongoing work...)
Growth curve model

\[ \hat{y}_{lmn} \rightarrow y_{lmn} \]

\[ K_{lm}, r_{lm}, \tau_{l}^{K}, \tau_{l}^{r}, \nu_{l} \]

Time Point

Repeat

orfΔ

Population

\[ K^{o}, r^{o}, \sigma^{K,o}, \sigma^{r,o}, \sigma^{r,p}, \sigma^{K,p}, \tau_{K,p}, \tau_{r,p}, P, \nu^{p} \]
Colony fitness

- The results of model fitting are estimates (or posterior distributions) of $r$ and $K$ for each yeast colony, and also the corresponding gene level parameters.
- Both $r$ and $K$ are indicative of colony fitness — keep separate where possible.
- Often useful to have a scalar measure of fitness — many possibilities, including $rK$, or MDR×MDP, where MDR is the maximal doubling rate and MDP is the maximal doubling potential.
- These statistical summaries can be fed in as data to the next level of analysis (or, ultimately, modelled jointly as a giant hierarchical model).
Epistasis

From Wikipedia:

- “Epistasis is the interaction between genes. Epistasis takes place when the action of one gene is modified by one or several other genes, which are sometimes called modifier genes. The gene whose phenotype is expressed is said to be epistatic, while the phenotype altered or suppressed is said to be hypostatic.”

- “Epistasis and genetic interaction refer to the same phenomenon; however, epistasis is widely used in population genetics and refers especially to the statistical properties of the phenomenon.”
Multiplicative model

- Consider two genes with alleles $a/A$ and $b/B$ with $a$ and $b$ representing “wild type” (note that $A$ and $B$ could potentially represent knock-outs of $a$ and $b$).
- Four genotypes: $aa$, $Ab$, $aB$, $AB$. Use $[\cdot]$ to denote some quantitative phenotypic measure (eg. “fitness”) for each genotype.

**Multiplicative model** of genetic independence:

$$[AB] \times [ab] = [Ab] \times [aB]$$  no epistasis

$$[AB] \times [ab] > [Ab] \times [aB]$$  synergistic epistasis

$$[AB] \times [ab] < [Ab] \times [aB]$$  antagonistic epistasis

- Perhaps simpler if re-written in terms of relative fitness:

$$\frac{[AB]}{[ab]} = \frac{[Ab]}{[ab]} \times \frac{[aB]}{[ab]}$$
Genetic independence and HTP data

- Suppose that we have scaled our data so that it is consistent with a multiplicative model — what do we expect to see?
- The independence model \([AB] \times [ab] = [Ab] \times [aB]\) translates to
  \[
  [\text{query} : abc\Delta] \times [\text{wt}] = [\text{query}] \times [abc\Delta]
  \]
- In other words
  \[
  [\text{query} : abc\Delta] = \frac{[\text{query}]}{[\text{wt}]} \times [abc\Delta]
  \]
- That is, the double-mutant differs from the single-deletion by a constant multiplicative factor that is independent of the particular single-deletion.
- ie. a scatter-plot of double against single will show them all lying along a straight line.
Assume that $F_{clm}$ is the fitness measurement for repeat $m$ of gene deletion $l$ in condition $c$ ($c = 1$ for the single deletion and $c = 2$ for the corresponding double-mutant).

Model:

$$F_{clm} \sim N(\hat{F}_{cl}, 1/\nu_{cl})$$

$$\log \hat{F}_{cl} = \alpha_c + Z_l + \delta_l \gamma_{cl}$$

$$\delta_l \sim Bern(p)$$

\(\delta_l\) is a variable selection indicator of genetic interaction

Then usual Bayesian hierarchical stuff…
Genetic interaction model

Stochastic modelling of genetic interaction in budding yeast
Genetic interaction results

![Graph showing the fitness (F) of orfΔ cdc13-1 double mutants at 27°C (doublings^2 / day) and orfΔ ura3Δ double mutants at 27°C (doublings^2 / day).]
Joint modelling of growth curves and genetic interaction

- We can integrate together the hierarchical growth curve model and the genetic interaction model into a combined joint model.
- This has usual advantages of properly borrowing strength, proper propagation of uncertainty, etc.
- Also very convenient to avoid requiring a scalar measure of “fitness”.
- If $y_{clmn}$ is the colony size at time point $n$ in repeat $m$ of gene $l$ in condition $c$, then

$$y_{clmn} \sim N(\hat{y}_{clmn}, 1/\nu_{cl})$$

$$\hat{y}_{clmn} = X(t_{clmn}; K_{cl}, r_{cl}, P)$$

$$\log K_{cl} \sim N(\alpha_c + K_{cl}^0 + \delta_l \gamma_{cl}, 1/\tau_{cl}K)$$

$$\log r_{cl} \sim N(\beta_c + r_{cl}^0 + \delta_l \omega_{cl}, 1/\tau_{cl}r)$$
Joint model

Stochastic modelling of genetic interaction in budding yeast
Joint model results
Summary

- Modern bioscience is generating large, complex data sets which require sophisticated modelling in order to answer questions of scientific interest.
- Big data forces trade-offs between statistical accuracy and computational tractability.
- Stochastic dynamic models are much more flexible than deterministic models, but come at a computational cost — the LNA can sometimes represent an excellent compromise.
- Notions of genetic interaction translate directly to statistical models of interaction.
- Big hierarchical variable selection models are useful in genomics.


