Design of a Motility Assay: The Statistical Point of View

Introduction

In this three-part tutorial we study the design and subsequent analysis of experiments such as motility assays. The problem is complex – very complex. We have little information on the statistical process involved. All we know is that there is a physiological process that underpins bacterium motility. It is likely that we will not be able to test comprehensive motility models against the data we will gather. On the positive side, we are not interested in building the best, most accurate model ever. What we really want is a model that is simple enough to be testable and that matches the data effectively.
In order to get an acceptable model, we are going to break the modelling process into the following three distinct steps – corresponding to the successive layers of a Bayesian data analysis routine.

Step 1: Generation of the hypotheses

Thanks to a bibliographical research, we can identify a set of hypotheses that we wish to compare to the data we have collected. In our case, a hypothesis consists in a model for the unhindered movement of bacteria.
Each candidate model should depend on a few parameters only – the present exercise is complex enough as it is.

Step 2: Match Each Hypothesis to the Data

Before knowing whether a model is supported by the data or not, we seek to match it to the data we have. In practice this step amounts to finding out what its best-fitting parameters are – see **Tutorial 2**.

Step 3: Hypothesis Testing

Once we have obtained for each model the best available match with the data, it is time to compare the various hypotheses. Again several approaches exist – see **Tutorial 2**.

Improving the Design of the Experiment

One of the most interesting properties of the Bayesian approach is that it provides intuitive methods for quantifying the accuracy of our predictions. As you will see the quantity and quality of the data available are crucial factors determining the quality of the whole process.
Being able to quantify the accuracy of our predictions has very appealing consequences. Of particular interest to us is the possibility to design our experiment so that the reliability and accuracy of the predictions improve.
Of course the design will be better if we have - and in particular if we have data representative of the phenomenon we wish to study. But interesting results can still be obtained with synthetic data generated by the various candidate models.

Overview of the Series of Tutorial

In the first tutorial we will focus on the construction of a relevant statistical model for the movement of a bacterium like b-sub and on the generation of the synthetic data required to train our analysis routines and design of the wetlab experiments

In the second part we will assume that we have an unrealistic level of control on the data acquisition process. Under such ideal assumptions, we will introduce the basics of Bayesian data analysis and how the results can be used to design experiments

Finally in the third tutorial, we will study the far more complex – and more realistic - case where the data acquisition process only gives us imperfect access to the data. As you will see the quality of the predictions that we can make is degraded and from an experimental point of view we need to gather more data.

Lessons to be learnt from the Tutorials

- When we know a little about a phenomenon, we can still effectively design experiments and effectively model the phenomenon
- It is not easy but it can be done!
- The Bayesian framework is ideal for the task at hand.

Recommendations

- Use Matlab for the calculations (since it is the iGEM sponsor and an industry standard)
- Try to solve the problem on paper first and then do simulations: you will learn a lot about the limitations of theory and the pitfalls of simulations….

One Last Word of Advice:

- Some of the questions are trick questions….

Tutorial 1: Creating Synthetic Data

Foreword on the Various Types of Data

There are 3 types of data:
- Synthetic Data
- Phantom Data
- Real Data

Bibliographical Research:
- What do these terms mean?
- When and why are they used?
- What are their respective advantages and limitations?

In our case we can only rely on synthetic data to conduct our pre-experimental analyses – including the development of relevant computer routines - and design our experiments.

[Discuss this statement]

Model Construction

Bibliographical Research on Bacterium Motility
- do a list of the relevant properties of the movement of b-subtilis
- turn them into an adequate model (or even better ***a family of models***).

To help you, here is a possible model.

[Discuss its relevance to the problem]

Running Phase:
Bacteria move at constant velocity v in direction Θ for a period of time T.
- v drawn from Gaussian of mean v_0 and standard deviation σ_1.
- T drawn from Gaussian of mean T_0 and standard deviation σ_2.
- Simulations: V_0 =10 ; T_0 =1s. Take $\sigma_1 = V_0/10$ $\sigma_2 = T_0/10$
- **Plot the distributions of v and T (or in (v,T))**

Rotating Phase:
Bacteria stop for a period of time T_s and rotate by an angle α.
- T_s drawn from Gaussian of mean T_r and standard deviation σ_r.
- α drawn from Von Mises of mean α_0 and parameter β.
- Simulations: α_0 =0 ; T_s =0.1s. Take β = 1 $\sigma_s = T_s/10$
- **Plot the distributions of α and T_s (or in (α, T_s))**

Simulations

What are the Data?

As we will see throughout this series of tutorials the choice of data to analyse is a crucial factor for the quality of the analysis. We therefore need to pause a little to think about what data we are likely to work with.

We can imagine that we will be able to shoot short movies of (a few) bacteria moving in a medium of our choice – that is within our control. **[Discuss]**

We can also imagine that these bacteria will have a few unhindered runs – that is without bumping into each other – within the field of view of the microscope. **[Discuss]**

Finally, we can imagine that with the right software the position of a few bacteria can be tracked with time. **[Discuss]**

Conclusion: Data = Position of bacteria well-chosen with time **[Discuss]**

Generation of Realistic/Plausible Synthetic Data

For each plausible model you have built:

- Simulate the run of a handful of bacteria over 5 minutes.
- If possible create a little movie (for Jamboree presentation)
- Store these data – they will be precious for the data analysis phase that will be developed in the next tutorial.

Generation of Unrealistic Synthetic Data

Whatever routine we develop must be capable of discriminating between models – given a set of available data. Furthermore, it is instructive to study the influence that bad data have on the quality (precision) of our predictions.

- Build another model with a running phase and a rotating phase. However, this time the velocity, time and angle distributions will be purposely unrealistic
- Simulate the run of a handful of bacteria over 5 minutes.
- Store these data – they will also be precious for the data analysis phase that will be developed in the next tutorial.

This unrealistic (wrong) model and the synthetic data you have generated with it will be very useful to test the routines you have developed – see **Tutorial 2**.

Qualitative Description of the motion of bacteria

In my opinion, one of the greatest challenges of Synthetic Biology will be the jump from the kind of qualitative description of experimental results that can be found in academic literature to qualitative results of the kind that are used in Engineering. This can be done for simple biobricks such as F2620 – and in general for simple biological phenomena. Whether this can be done in general is unclear…

To finish this tutorial, it is worth looking at the advantages of a quantitative description. To do this, Let us try to find out when a qualitative description is enough and when it needs complementing. Synthetic data are ideal to conduct this kind of study since we control everything in the study and know the underlying truth.
So let us consider one of your realistic models and the wrong model you have just created.

- Give a qualitative description of the motions generated by both models
- Make parameters of your 'wrong' model vary to the point when it becomes hard to tell the motion generated by your 'realistic' model from the motion generated by the other model
- Discuss

NB: In the second and third tutorials we will see that obtaining a quantitative description is not that easy and often a certain amount of uncertainty remains at the end of the analysis.

At the End of this Tutorial you should…

- Understand the differences between the various classes of data
- Be able to draw from the standard statistical distributions with Matlab
- Build a sensible model for the random walk of a bacterium
- Simulate it with Matlab and store it in an effective manner