Data Analysis

Richard Sear

University of Surrey

RAMP (MSCA ITN) School, Ireland May 2018

Outline

1. We all need models . . . to make predictions
2. Linear regression & two types of errors in fitting
3. The problem of sampling high dimensional space – finding a needle in multidimensional haystack
4. Partial solutions to searching high dimensional space: 1) get more data, 2) analyse data better, including scoring outcomes & image analysis — discussion

Prediction

Prediction

In science we want to predict things, eg what solution conditions should I try to crystallise my membrane protein, in other words: What solution conditions have the highest probability of yielding diffraction quality crystals?

Membrane protein systems are complex so it is hard to answer this question, but even a very approximate answer is better than nothing.

Prediction almost always uses a model. Here, a model is defined as a simplified representation of the system (eg membrane-protein solution) of interest. When you fit, eg a straight line to data, that is a model. You are making the simplifying assumption that one variable (y) varies linearly with another one (x).

Models

"Essentially, all models are wrong, but some are useful" – George Box

"Now it would be very remarkable if any system existing in the real world could be exactly represented by any simple model. However, cunningly chosen parsimonious models often do provide remarkably useful approximations. For example, the law PV = RT relating pressure P, volume V and temperature T of an "ideal" gas via a constant R is not exactly true for any real gas, but it frequently provides a useful approximation and furthermore its structure is informative since it springs from a physical view of the behavior of gas molecules." — George Box (statistician)

Linear regression

Fitting & Errors

Errors when comparing data & model

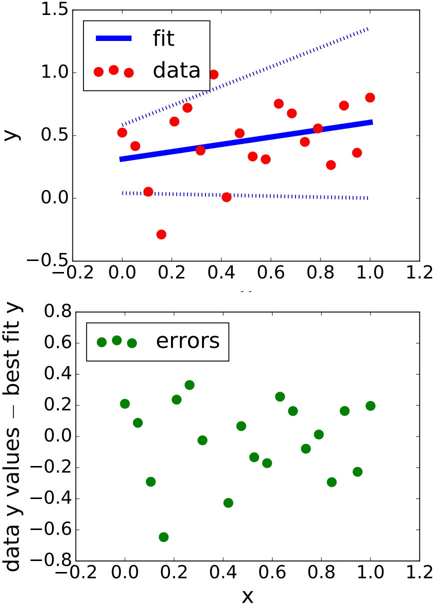
1. Model is wrong
2. Data is noisy

2. is statistical error, would be eliminated with enough data, whereas 1. remains however much data you have. Distinguishing 1. and 2. can be difficult if data is limited/noisy.

Linear regression: Data is noisy

Model, here assumption that y varies linearly with x: $y = mx + c$, is right but data is noisy.

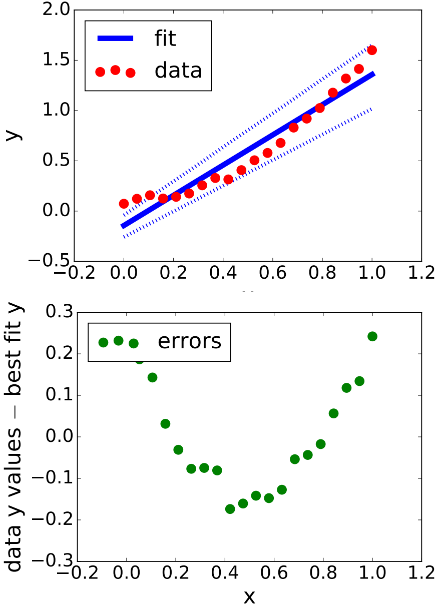
Note that errors between fit and points has no trends — errors are random!



Linear regression: Model is wrong

Model, here assumption that y varies linearly with x: $y = mx + c$, is wrong.

Note that errors between fit and points have clear trends — errors are not random!



Complexity & many parameters

Will now look at sampling many parameters — salt, polymer etc concentrations in the case of membrane-protein crystallisation

Simple things are easy to predict, complex things are hard to predict

Simple question: When is the next lunar eclipse in Dublin?

Answer: 9:30 pm Friday 27th July 2018

For eclipse, only consider two bodies' (Sun & Moon) positions relative to Earth

Under what conditions will membrane protein X crystallise?

Answer: I don't know

Many parameters ....

NB For complex systems, need all the data we can get. In context of membrane protein crystallisation, this means data from all trials not just successful ones. Unsuccessful trials give useful data.

Sampling many parameters

General proteins: Rupp, J Struct Biol **142**, 162 (2003).

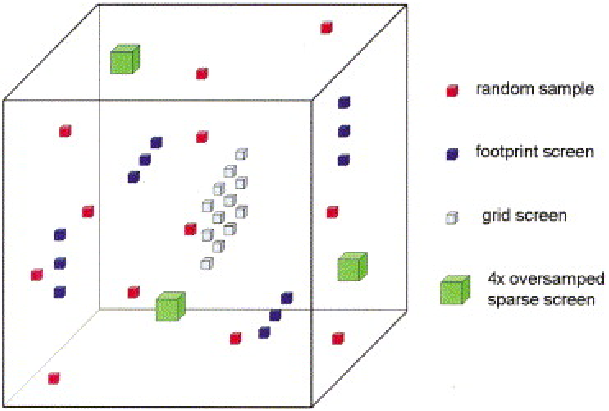
Membrane proteins: Asur et al., Bioinformatics **22**, e40 (2006)

Need to record unsuccessful as well as successful trials: Newman et al., Acta Cryst. F **68**, 253 (2012)

Outcome of crystal trial classification scheme: Luft et al., Cryst. Growth Design **11**, 651 (2011)

NB Even if you take many parameters into account, there is no guarantee that you have captured all relevant parameters, eg if mixing method affects crystal nucleation then if this is not taken into account ....

Sampling and crystallisation



Schematic from Rupp (J Struct Biol, 2003). Showing sampling three dimensions of parameter space, axes could be concentration of, eg, NaCl, 1000 MW PEO, ... Note, only plotting three dimensions but in reality many more than three salts, polymers, etc to vary.

The Curse of Dimensionality

If there is one variable, eg can only vary concentration of one variable, eg NaCl, then rough idea of behaviour is obtained from ≈ 10 NaCl concentrations.

If there two salts, say NaCl and LiCl, need $10^2 = 100$ points.

For d possible salts, polymers etc, whose concentration you can vary, need $\approx 10^d$ data points, $= 100,000,000,000,000,000,000$ points when $d = 10$.

This is impossible! — the Curse of Dimensionality

Caveats, don't need complete dependence of crystallisation on d variables:

1. Especially for large d, not all d variables independent, roughly speaking if say LiCl and NaCl have similar effects then the can replace pair $(c_{LiCl}, c_{NaCl}$ with $c_{LiCl} + c_{NaCl}$, effectively reducing d by one. Maths for this: Principle Component Analysis.
2. Just want good enough crystals ...

Possible (partial) solutions to the problem of searching this huge parameter space

Jen’s email RE: Fri am discussion

We will have a discussion centred on the topic of “Rationalising Membrane Protein Crystallization”. To keep this focussed, we would like you to prepare one (maximum two) powerpoint slides on your view of how we can rationalise membrane protein crystallization. This could take the form of an idea or approach that you take in your research, which you can convince us is a good way to proceed, or a question that you would like to pose to the group of ESRs (e.g. if you are looking for a collaborator or help with a specific aspect of your project). The hope is that the discussion will emphasise some of the key questions we are trying to answer through this network project. The discussion will be informal and interactive.

Standing on the shoulders of giants: Others have searched this huge space

Others have searched this space, but for their proteins, and found crystals. Can we take advantage of their work?

If your protein, B, is similar in some sense to that of an already crystallised protein, A, try the conditions that worked for A.

Another way of looking at this: if the fraction of the *d*-dimensional parameter space where a trial yields a useful result is very small, but the correlation between one membrane protein and another is high, than it is rational to just try for membrane B, what worked for membrane protein A.

Partial solutions to the problem of searching huge parameter space

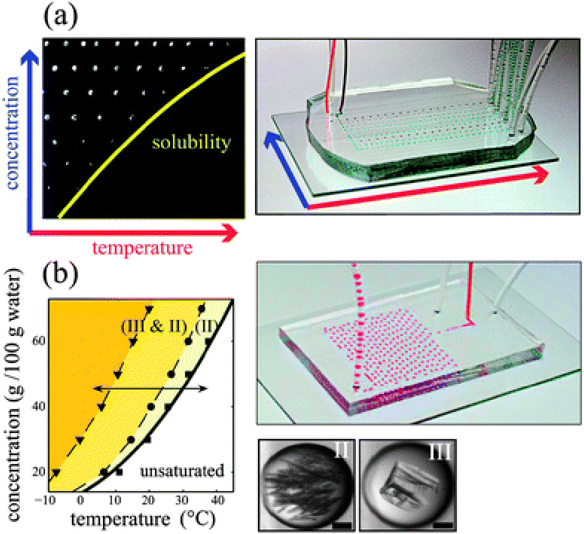
1. Do more trials: microfluidics to do more trials, faster and with less protein. Automated image analysis to get trial results faster.
2. Get more information from same number of trials: Extract info from trials, eg crystal size, amorphous or crystal, XRD peak width, ... and look for correlations between outputs (eg presence of crystal) and inputs (eg, conc of precipitant X).

For 2:The greatest teacher, failure is — *Yoda* (The Last Jedi).
Yoda is right, we can learn from trials that did not result in crystals.
But published work just records conditions that worked, not the much more numerous conditions that did not. Newman et al., Acta Cryst. F **68**, 253 (2012) clearly stated this but appear not to have acted on their recommendations.

More trials faster

Microfluidics – work with μl, nl, pl, or fl volumes, droplets μm to mm across.

Do many (100s +) trials in one experiment, with small amount



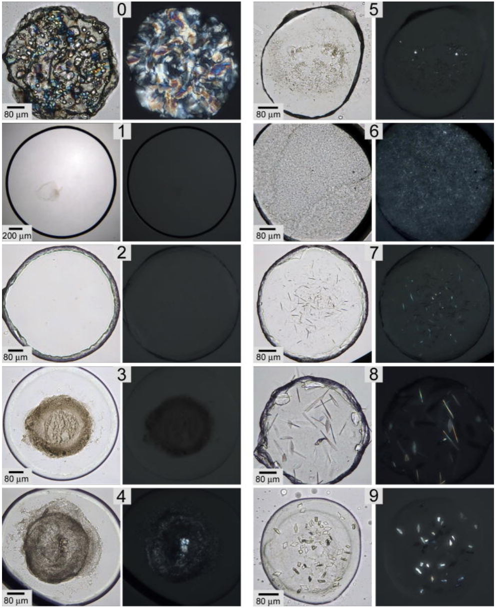
Leng & Salmon, Lab on Chip **9**, 24 (2009)

Analysing trial results

1. Scoring trial outcomes
2. Analysing images: feature detection, machine learning
3. Analyse features (eg peak width) of XRD patterns — I know nothing about this but see my lecture on crystals for background

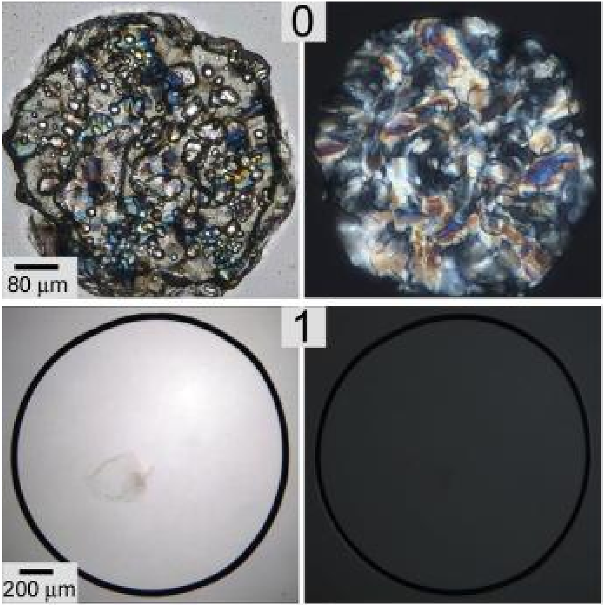
Membrane protein crystallisation: Scoring outcomes

Scoring trial outcomes



Scoring the outcome of in meso crystallization trials. The scale runs from 0 to 9 and the corresponding images recorded in normal light (left panel) and between crossed polarizers (right panel) are shown. Caffrey et al. Nature Protocol **4** 706 (2009)

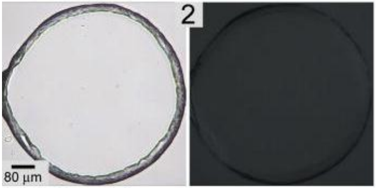
Scoring trial outcomes

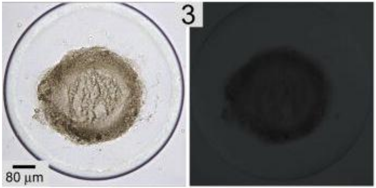


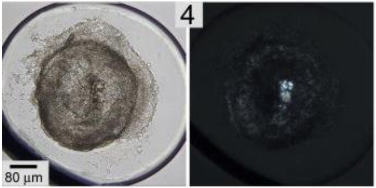
0 – Birefringent mesophase. Lamellar and hexagonal phases are birefringent under cross-polarizers.

1 – Dissolved lipidic mesophase.

Scoring trial outcomes





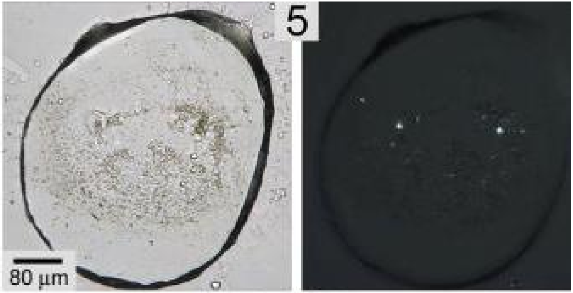


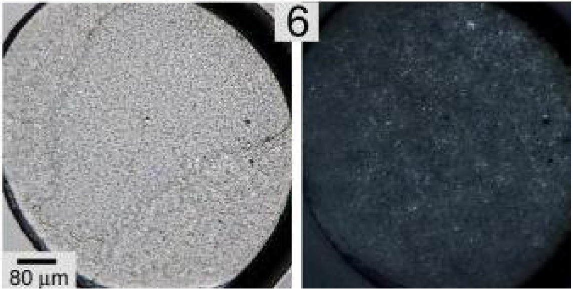
2 – Clear cubic phase. Non-birefringent, transparent, and gel-like with rough edges.

3 – Precipitate. Protein forms a brownish, non-birefringent precipitate.

4 – Birefringent precipitate.

Scoring trial outcomes

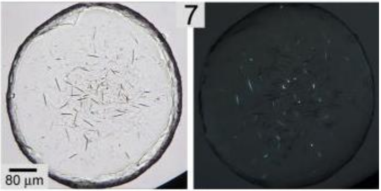


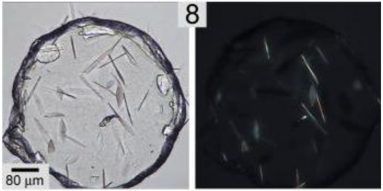


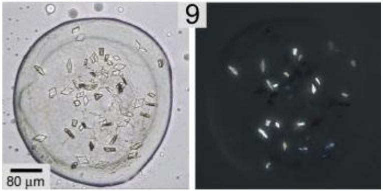
5 – Crystallites or spherulites. Protein forms birefringent particles that lack angularity or a well-defined crystal shape.

6 – Microcrystals. ‘Shower’of crystals $< 1\mu$.

Scoring trial outcomes







7 – Needles. Crystals grow preferentially in 1 dimension.

8 – 2D Plates. Crystals grow in two dimensions.

9 – 3D Crystal.

Image analysis

1. Feature recognition and quantification — we will do a bit of this
2. Machine learning approach (see Bruno et al, arXiv:1803.10342 (2018)) use algorithms that you give 1000s of scored (by humans) images of crystallisation trials, and then learn from these images how to score outcomes themselves (without further human intervention).