Introduction to maximum likelihood

Airlie McCoy

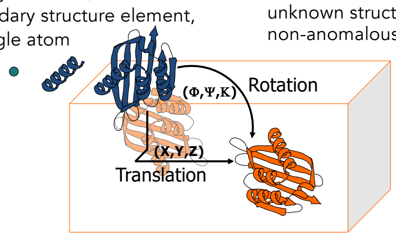
UNIVERSITY OF CAMBRIDGE

1

molecular replacement

quaternary assembly,
tertiary structure,
secondary structure element,
or single atom

unknown structure,
non-anomalous data



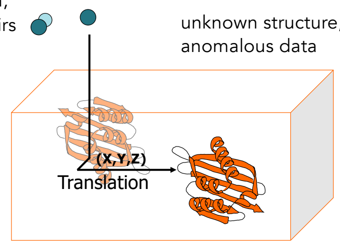
2

2

single anomalous dispersion phasing

single atom,
or atom pairs

unknown structure,
anomalous data



3

3

software for phasing

Score: find the pose that best fits the data
Search: strategies for exploring poses

4

4

software for molecular replacement

- MolRep
- AMoRe
- CNS

Patterson RF/CC TF
Amplitudes
RF+TF

- EPMR
- COMO
- SOMoRe
- Queen of Spades

CC (TF)
Amplitudes
6D (but not exhaustive)

- Phaser

Maximum likelihood
Intensities
RF+TF+RF+TF

5

5

log-likelihood gain on intensity

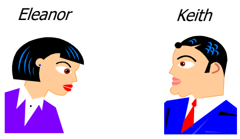
$$LLGI = \sum_{hkl\,info} log\left[\frac{2E_e}{1-D_{obs}^2\sigma_A^2} exp\left(-\frac{E_e^2 + D_{obs}^2\sigma_A^2 E_C^2}{1-D_{obs}^2\sigma_A^2}\right) I_0\left(\frac{2E_e D_{obs}\sigma_A E_C}{1-D_{obs}^2\sigma_A^2}\right)\right]$$

6

6

Concepts:

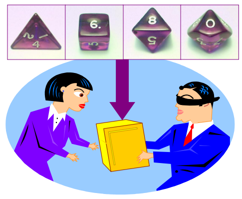
- Maximum Likelihood
- Independence
- Log(Likelihood)
- Bayes' Theorem
- Integrating out Variables
- Central limit theorem



7

7

a game of dice

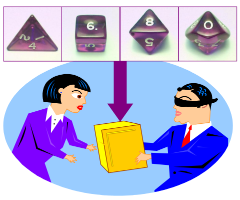


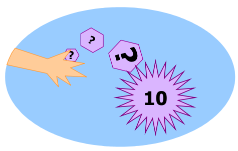
- Put four unbiased dice in a box
- Keith selects a die at random
- How often will Keith guess correctly which die he selected?

8

8

a game of dice with data





- Put four unbiased dice in a box
- Keith selects a die at random
- Eleanor rolls the die and tells Keith the result of the roll
- Which die does Keith guess he selected?
- How confident will he be that he is right?

9

9

maximum likelihood

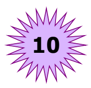
- Roll a 10
- The die obviously must have been the 10 sided die
- What does "must" mean in probabilities?

$P(10;\boxed{10}) = \frac{1}{10}$

$P(10;\boxed{8}) = 0$

$P(10;\boxed{6}) = 0$

$P(10;\boxed{4}) = 0$



- 10-sided die is most likely

10

10

maximum likelihood

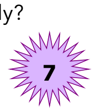
- Roll a 7
- The die could have been
  - the 10 sided or
  - the 8 sided die
- Which die is most likely?

$P(7;\boxed{10}) = \frac{1}{10}$

$P(7;\boxed{8}) = \frac{1}{8}$

$P(7;\boxed{6}) = 0$

$P(7;\boxed{4}) = 0$



- 8-sided die is most likely

11

11

maximum likelihood

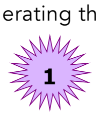
- Roll a 1
- Could have been rolled by any of the dice
- The most likely die is the one with the highest probability of generating the data

$P(1;\boxed{10}) = \frac{1}{10}$

$P(1;\boxed{8}) = \frac{1}{8}$

$P(1;\boxed{6}) = \frac{1}{6}$

$P(1;\boxed{4}) = \frac{1}{4}$



- 4-sided die is most likely

12

12

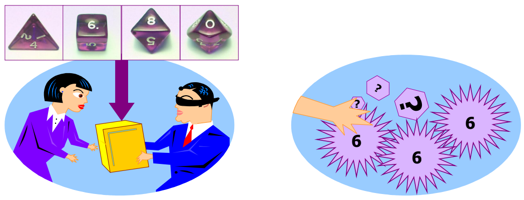
maximum likelihood: crystallography

- Data are in reciprocal space
  - Intensities and sigmas of the intensities
- Model is the structure in real space
  - Calculate the structure factor F_C from the model to put the model information into reciprocal space
- Compare F_C with observed data I_{obs} and σ_{Iobs}
- "Solution" is the model with the highest likelihood of generating I_{obs}

13

13

a game of dice with more data



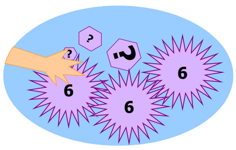
- Put four unbiased dice in a box
- Keith selects a die at random
- Eleanor rolls that die three times and tells Keith the results
- Which die does Keith guess he selected?
- How confident will he be that he is right?

14

14

a game of dice with more data

- What is the chance of throwing a 6 three times from a 6-sided die?
- The chance of throwing a 6, or any other number, the second, or third time is not influenced by the value of the first roll - they are independent
- When probabilities are independent, they multiply



15

15

independence and log-likelihood

$$P(6;\boxed{6}) = \tfrac{1}{6} = 0.16666667$$

$$P(6,6;\boxed{6}) = \tfrac{1}{6} \times \tfrac{1}{6} = \tfrac{1}{36} = 0.0277778$$

$$P(6,6,6;\boxed{6}) = \tfrac{1}{6} \times \tfrac{1}{6} \times \tfrac{1}{6} = \tfrac{1}{216} = 0.0046296$$

$$P(6\ldots\times 100;\boxed{6}) = 6^{-100} = 1.53064 \times 10^{-78}$$

16

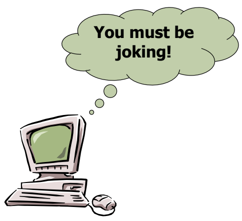
16

independence and log-likelihood

"Oh great one, what is the probability of throwing a 6 from a six-sided die one billion times?"

```
> SYSTEM-F-FLTOVF_F, arithmetic fault,
floating underflow at PC=00006244,
PSL=03C0 0020 %TRACE-F-TRACEBACK,
symbolic stack dump follows
module name      routine name    line
OVERF            OVERF           104
DPARA$MAIN       DPARA$MAIN      276
```

- Computers can not store numbers very close to zero



17

17

independence and log-likelihood

"Oh great one, what is the logarithm of the probability of throwing a 6 from a six-sided die one billion times?"

```
> -778151250.4
```

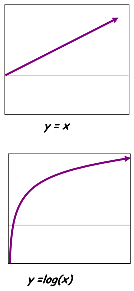
- log(likelihood) is not close to zero
  - So the log(likelihood) solves the small number problem
- but can we just switch to using the log(likelihood)?

18

18

independence and log-likelihood

- Logarithmic functions are "monotonic" functions
  - i.e. they "preserve the given order"
  - If y1<y2 for all x1<x2 then log(x1)<log(x2)
- The parameter values obtained optimising log(likelihood) are the same as those obtained optimising likelihood
  - Optimising log(likelihood) ≡ Optimising likelihood

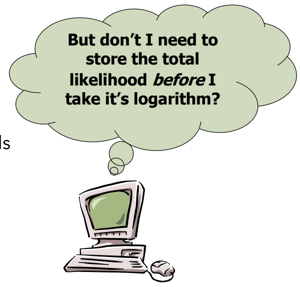


19

19

independence and log-likelihood

- No, there is a shortcut to the log(total likelihood) when total likelihood is a product of likelihoods
- If log(total likelihood) equals log(Π likelihoods)
- Then log(total likelihood) also equals Σlog(likelihoods)



20

20

independence and log-likelihood

log(Πlikelihoods) = Σlog(likelihoods)

$$\begin{aligned}&\log\left(P\left(3,3;\boxed{6}\right)\right)\\&=\log\left(P\left(3;\boxed{6}\right)\times P\left(3;\boxed{6}\right)\right)\\&=\log\left(\tfrac{1}{6}\times\tfrac{1}{6}\right)\\&=\log(0.0277)\\&=-1.556\end{aligned}$$

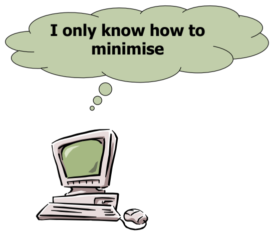
$$\begin{aligned}&\log\left(P\left(3,3;\boxed{6}\right)\right)\\&=\log\left(P\left(3;\boxed{6}\right)\right)+\log\left(P\left(3;\boxed{6}\right)\right)\\&=\log\left(\tfrac{1}{6}\right)+\log\left(\tfrac{1}{6}\right)\\&=-0.778-0.778\\&=-1.556\end{aligned}$$

21

21

independence and log-likelihood

- Computer algorithms are designed to minimise
- Therefore we optimise our parameters by minimising the -log(likelihood)



22

22

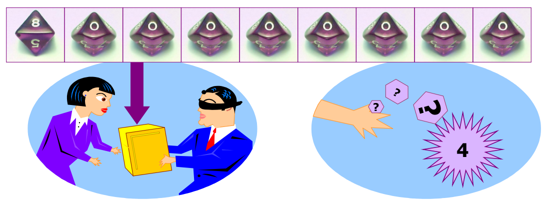
independence and log-likelihood

- In Phaser, we assume reflections are independent
- This is not true!
  - broken by solvent and non-crystallographic symmetry
  - correlations essential for
    - direct methods
    - solvent flattening, non-crystallographic symmetry averaging
- Total likelihood is the product of the likelihoods for each reflection
  - Algorithms calculate the sum of reflection log(likelihood)
- Maximum likelihood search and refinement algorithms minimise the -log(likelihood)

23

23

a game of dice with copies of a die



- Put one 8-sided die and eight 10-sided dice in a box
- Keith selects a die at random
- Eleanor rolls the die and tells Keith the result of the roll
- Which die does Keith guess he selected?

24

24

prior probability and Bayes' theorem

- Bayes' Theorem

$$P(\text{model};\text{data}) = \frac{P(\text{model})}{P(\text{data})} \times P(\text{data};\text{model})$$

- In experimental situations, P(data) is constant, and when comparing probabilities can be ignored

$$P(\text{model};\text{data}) = \underbrace{P(\text{model})}_{\text{prior probability}} \times \underbrace{P(\text{data};\text{model})}_{\text{likelihood}}$$

25

25

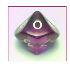
Bayes' theorem

- Roll a 4
- The prior probability of selecting the 10-sided die dominates the higher likelihood of throwing a 4 from the 8-sided die than from the 10-sided die

$$P(\boxed{10};4) = P(\boxed{10})P(4;\boxed{10})$$
$$= \tfrac{8}{9} \times \tfrac{1}{10}$$
$$= \tfrac{8}{90}$$
$$= 0.0888$$

$$P(\boxed{8};4) = P(\boxed{8})P(4;\boxed{8})$$
$$= \tfrac{1}{9} \times \tfrac{1}{8}$$
$$= \tfrac{1}{72}$$
$$= 0.01388$$

- 10-sided die is most likely



26

26

Bayes' theorem: crystallography

- Bayes' Theorem is used in refinement

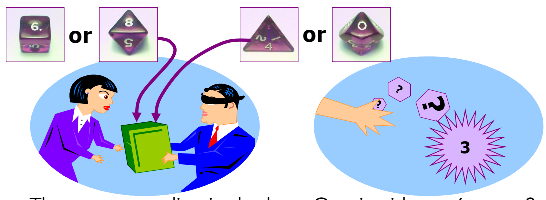
$$P(\text{model;data}) = P(\text{model}) \times P(\text{data;model})$$

  - Prior probability from the chemistry i.e. knowledge of bond-lengths, bond-angles, planarity etc
  - Likelihood from the X-ray diffraction experiment
- Also used in density modification
- Fundamental principle in the method of "integrating out nuisance variables"…

27

27

a game of dice with unknown dice



- There are two dice in the box. One is either a 6 or an 8 sided die and the other is either a 4 or a 10 sided die
- Keith selects a die at random
- Eleanor rolls, and tell Keith the result
- Which die does Keith guess he selected?

28

28

integrating out nuisance variables

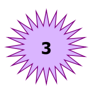
- We consider both possibilities (for each box) by <u>summing</u> the probabilities of each possible die

29

29

integrating out nuisance variables

$$P(\boxed{6 \text{ or } 8};3) = P(\boxed{6};3) + P(\boxed{8};3)$$
$$= P(3;\boxed{6}) \times P(\boxed{6}) + P(3;\boxed{8}) \times P(\boxed{8})$$
$$= \left(\tfrac{1}{6} \times \tfrac{1}{2}\right) + \left(\tfrac{1}{8} \times \tfrac{1}{2}\right)$$
$$= 0.1458\overline{3}$$



$$P(\boxed{4 \text{ or } 10};3) = P(\boxed{4};3) + P(\boxed{10};3)$$
$$= P(3;\boxed{4}) \times P(\boxed{4}) + P(3;\boxed{10}) \times P(\boxed{10})$$
$$= \left(\tfrac{1}{4} \times \tfrac{1}{2}\right) + \left(\tfrac{1}{10} \times \tfrac{1}{2}\right)$$
$$= 0.175 \quad \textit{most likely}$$

30

30

integrating out nuisance variables

- Probability for discrete probabilities

$$P(\text{data};\text{model}) = \sum_{i=1}^{n} P(\text{data},x_i;\text{model}), \quad \text{where } a \le x_i \le b$$

- For continuous probability, sum becomes an integral

$$P(\text{data};\text{model}) = \int_a^b P(\text{data},x;\text{model})\,dx$$

- The unknown variable is called a nuisance variable
  - The removal of a nuisance variable from a probability distribution by integration is called "integrating out" the nuisance variable
- Nuisance variables can be very useful!

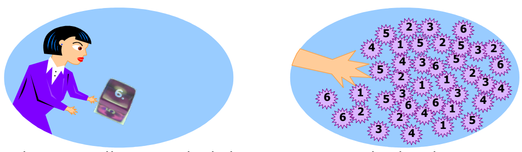
31

the phase problem

- Data (for each reflection) is the observed intensity
  - This is converted into a structure factor amplitude $|F_O|$
- Model is the phased calculated structure factor F_C
- Clever bit: Probabilities are calculated in terms of the phased observed structure factor F_O
  - (and the calculated structure factor F_C)
- The introduced phase of F_O is a nuisance variable
- Probability of $|F_O|$ is then found by integrating out the "nuisance" (but very useful) phase

32

the average of many games of dice



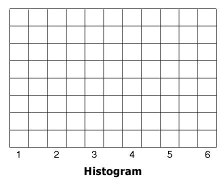
- Eleanor rolls a 6-sided die 40 times and take the average of the values
- Eleanor does this many times, plotting the average values from each game on a histogram
- What is the final distribution of values in the histogram?

33

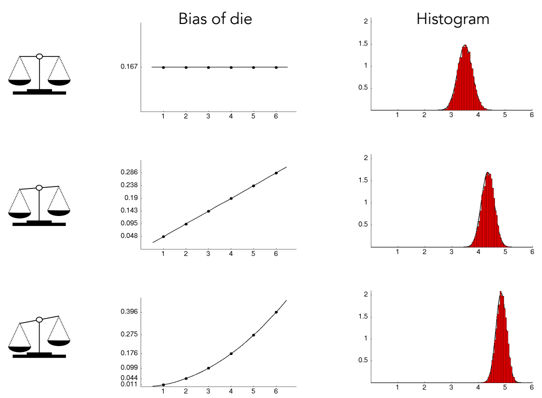
the average of many games of dice

- Build up a histogram of the average 40 throws of the dice

1. $40\times\boxed{3}$
2. $20\times\boxed{3}+20\times\boxed{4}$
3. $10\times\boxed{1}+10\times\boxed{2}+10\times\boxed{5}+10\times\boxed{6}$
4. $40\times\boxed{4}$
5. $10\times\boxed{1}+10\times\boxed{2}+10\times\boxed{3}+10\times\boxed{4}$
6. $10\times\boxed{2}+10\times\boxed{3}+10\times\boxed{4}+10\times\boxed{5}$
7. $20\times\boxed{4}+20\times\boxed{5}$
8. $10\times\boxed{3}+10\times\boxed{5}+20\times\boxed{4}$
9. $10\times\boxed{1}+10\times\boxed{5}+20\times\boxed{3}$
10. $20\times\boxed{2}+20\times\boxed{5}$



34



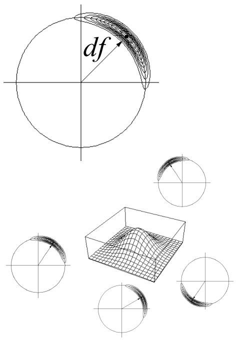
35

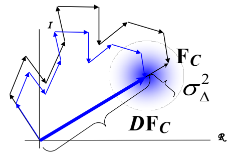
central limit theorem

- No matter what the bias of the die, the histogram generated by the average of many rolls of the die is a Gaussian
- This is true even when the bias if the die (from which the average is computed) is decidedly not Gaussian
- This property is called the central limit theorem
- Historically, the central limit theorem was called the "law of errors"

36

structure factor for a model with errors



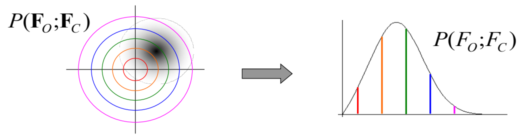


$$P(\mathbf{F}_O;\mathbf{F}_C)=\frac{1}{\pi\sigma_\Delta^2}e^{-\frac{|\mathbf{F}_O-D\mathbf{F}_C|^2}{\sigma_\Delta^2}}$$

37

likelihood for a model with errors

- But we do not measure the phase of the observed structure factor!
- Integrate out the phase to get the likelihood for the unphased structure factors



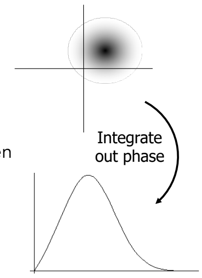
$$P(F_O;F_C)=\frac{2F_O}{\sigma_\Delta^2}e^{-\frac{F_O^2+D^2F_C^2}{\sigma_\Delta^2}}I_0\left(\frac{2F_ODF_C}{\sigma_\Delta^2}\right)$$

38

likelihood for a model with errors

$$P(F_o;F_c)=\frac{2F_o}{\sigma_\Delta^2}exp\left(-\frac{F_o^2+D^2F_C^2}{\sigma_\Delta^2}\right)I_0\left(\frac{2F_oDF_C}{\sigma_\Delta^2}\right)$$

"2D Gaussian distribution"



"Rice distribution"

- Express in terms of normalized structure factors – E values
  - Average value of intensity at any given resolution shell is 1

$$P(E_o;E_c)=\frac{2E_o}{1-\sigma_A^2}exp\left(-\frac{E_o^2+\sigma_A^2E_C^2}{1-\sigma_A^2}\right)I_0\left(\frac{2E_o\sigma_AE_c}{1-\sigma_A^2}\right)$$

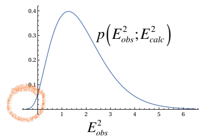
I_0: modified Bessel function of order 0

39

39

maximum likelihood scoring

- Peak intensity minus background can yield negative net intensity
- I_{obs} and σ_{Iobs} are used to generate new terms:
  - Effective E "E_e" and "D_{obs}"
- E_e replaces E_{obs} and σ_A is inflated with D_{obs} term in the Rice equation



Exact intensity likelihood function for $E_C = 1.5$, $\sigma_A = 0.8$ and $\sigma_{obs} = 0.2$

$$LLGI=\frac{2E_e}{1-D_{obs}^2\sigma_A^2}exp\left(-\frac{E_e^2+D_{obs}^2\sigma_A^2E_C^2}{1-D_{obs}^2\sigma_A^2}\right)I_0\left(\frac{2E_eD_{obs}\sigma_AE_C}{1-D_{obs}^2\sigma_A^2}\right)$$

40

40

information content and speed

$$LLGI=\sum_{hkl}log\left[\frac{2E_e}{1-D_{obs}^2\sigma_A^2}exp\left(-\frac{E_e^2+D_{obs}^2\sigma_A^2E_C^2}{1-D_{obs}^2\sigma_A^2}\right)I_0\left(\frac{2E_eD_{obs}\sigma_AE_C}{1-D_{obs}^2\sigma_A^2}\right)\right]$$

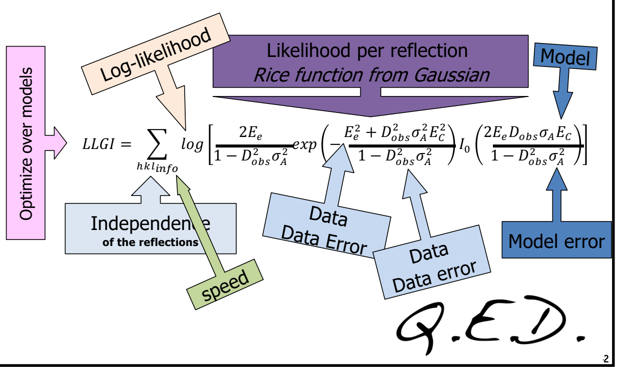
- The 'information content per reflection' allows data to be individually assessed and rejected if the information content is low
- Sum over fewer reflections without losing signal
- There is no penalty for inputting data beyond traditional resolution limits or I on sigI limits

$$LLGI=\sum_{hklinfo}log\left[\frac{2E_e}{1-D_{obs}^2\sigma_A^2}exp\left(-\frac{E_e^2+D_{obs}^2\sigma_A^2E_C^2}{1-D_{obs}^2\sigma_A^2}\right)I_0\left(\frac{2E_eD_{obs}\sigma_AE_C}{1-D_{obs}^2\sigma_A^2}\right)\right]$$

41

41

log-likelihood gain on intensity



$$LLGI=\sum_{hklinfo}log\left[\frac{2E_e}{1-D_{obs}^2\sigma_A^2}exp\left(-\frac{E_e^2+D_{obs}^2\sigma_A^2E_C^2}{1-D_{obs}^2\sigma_A^2}\right)I_0\left(\frac{2E_eD_{obs}\sigma_AE_C}{1-D_{obs}^2\sigma_A^2}\right)\right]$$

2

42

what can we learn from the equation?

- Intensities and their errors are the best for accurately representing the experimental data
  - not French and Wilson truncated structure factors
- Weak data cannot be simply excluded with anisotropic resolution limit when there is anisotropy
  - Truncation perturbs normalization to E-values
  - Weak intensities are important for data correction
- Each reflection adds precious information
  - Every reflection counts, and this is regardless of resolution
- Accurate estimates of the errors are very important
  - Better errors, more signal

43

43