Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Liking likelihood

Airlie J. McCoy

University of Cambridge, Department of Haematology, Cambridge Institute for Medical Research, Wellcome Trust/MRC Building, Hills Road, Cambridge CB2 2XY, England

Correspondence e-mail: ajm201@cam.ac.uk

Received 26 January 2004
Accepted 30 June 2004

Maximum-likelihood methods have now been applied to most areas of macromolecular crystallography, including data reduction, molecular replacement, experimental phasing and refinement. However, students of macromolecular crystallography are predominantly taught only traditional crystallographic methods and therefore have little understanding of the methods underlying the modern software that they routinely use in structure determination. This situation arises, at least in part, because maximum likelihood is considered to be too difficult to be taught to students who lack substantial mathematical training within the limited time frame of undergraduate/graduate courses. A method of introducing maximum-likelihood concepts with the help of dice is described here and it is then shown how these concepts can form the core of understanding maximum-likelihood refinement, molecular replacement and experimental phasing. Within the framework described, the crystallographic maximum-likelihood techniques are all reduced to the same basic concepts and become easier and less time-consuming to teach than traditional methods, which rely on disparate concepts.

1. Introduction

Maximum likelihood is a branch of statistical inference that asserts that the best hypothesis (*i.e.* set of parameters, which includes estimates of the errors) on the evidence of the data is the one that explains what has in fact been observed with the highest probability. In the context of macromolecular crystallography, maximum likelihood has come to refer to the set of new statistical methods that improved upon the least-squares methods that preceded them. The least-squares methods were not contrary to the principle of maximum likelihood, since least squares is a special case of maximum likelihood where the errors in the parameters are simple Gaussians, rather than more complex functions. The slow acceptance of maximum likelihood was therefore not because maximum likelihood itself was considered inappropriate, but because least squares works acceptably when the data and model are good and because computers were not capable of performing the more complex calculations required for more sophisticated maximum-likelihood treatments in reasonable times. Maximum likelihood is not the only method for obtaining a set of parameters from experimental data. Indeed, in other fields of application maximum likelihood may not be the best method, as maximum-likelihood estimators can be severely biased. However, maximum likelihood gives little bias when applied in crystallography and it has been extremely successful in supplying better probability models, particularly

when the data and/or model are poor, and has been instrumental in the solution of numerous macromolecular structures.

Dice have a long history in the explanation of problems in likelihood, maximum entropy and Bayesian theory (*e.g.* Jaynes, 1968, 1979; Frieden, 1985; Mohammed-Djafari, 2003). In this tradition, I present here basic maximum-likelihood concepts using thought experiments with dice. These concepts are then used to explain maximum-likelihood refinement, molecular replacement and experimental phasing.

2. Experiments with dice

There are six important concepts that are needed in order to understand the statistical approach of maximum likelihood in crystallography: maximum likelihood, independence, log-likelihood, Bayes' theorem, integrating out nuisance variables and the central limit theorem. These concepts will be explored with the help of dice with different numbers of sides (Fig. 1).

2.1. Dice and probability

A game of dice.

I put four unbiased dice in a box: one four-sided, one six-sided, one eight-sided and one ten-sided.

I select a die at random.

How often will you guess correctly which die I selected?

It is obvious that there is a one in four chance of getting the correct answer. If the experiment is performed a large number of times you will guess the answer a quarter of the time, or if a large number of people guess each time a quarter will guess correctly.

2.2. Dice and maximum likelihood

A game of dice with data.

I put four unbiased dice in a box: one four-sided, one six-sided, one eight-sided and one ten-sided.

I select a die at random.

I roll the die and tell you the result of the roll.

Which die was the most likely to be selected?

If I were to roll a 10, it is obvious that the die selected must have been the ten-sided die. Why is it obvious? Because the probability of rolling a 10 from the four-, six- or eight-sided die is zero, but the probability of rolling a 10 from the ten-sided die is non-zero. The probability is written as

$$P(10; \mathbf{10}) = \tfrac{1}{10},$$

where the semi-colon means 'given' (for a glossary of terms see Table 1) and I have denoted the type of die by its number of sides in bold. The probability of the observed data (the number rolled) given the model (the number of sides of the die) is called the likelihood.

What would be the case if I rolled a 7? If the same analysis is performed again, the likelihood of rolling a 7 from the four- or six-sided die is 0, but the likelihood of rolling a 7 from the eight-sided die is one in eight and the likelihood of rolling a 7 from the ten-sided die is one in ten. Therefore, it is most likely that the eight-sided die would have been selected. What if I roll a 1? It is most likely that the four-sided die would have been selected. The most likely die is the one with the highest likelihood of generating the data: this is the principle of maximum likelihood.

How confident are you that the die is an eight-sided die if the roll was a 7? Not very, because the difference between the likelihood of rolling a 7 from the eight-sided and ten-sided die is only small. The ratio between two likelihoods is a measure of confidence (known as the likelihood ratio). For example, when I roll a 10, the likelihood ratio agrees that you are supremely confident that I selected a ten-sided die, rather than, say, the eight-sided die,

$$\frac{P(10; \mathbf{10})}{P(10; \mathbf{8})} = \frac{\frac{1}{10}}{0} = \infty.$$

In the case where I roll a 7, the likelihood-ratio is close to 1 (the ratio for equal likelihoods),

$$\frac{P(7; \mathbf{8})}{P(7; \mathbf{10})} = \frac{\frac{1}{8}}{\frac{1}{10}} = 1.25.$$

Table 1
Glossary of terms.

$\sum_{i=1}^{N} a_i = a_1 + \ldots + a_N$	Sum of all a_i for i between 1 and N
$\prod_{i=1}^{N} a_i = a_1 \times \ldots \times a_N$	Product of all a_i for i between 1 and N
$\{a_i\} = a_1, \ldots, a_N$	Set of all a_i
$\int_a^b f(x)\,\mathrm{d}x$	Definite integral of function $f(x)$ between values a and b
$I_0(x)$	Modified Bessel function of order 0 with argument x
$P(A; B)$	Probability of A, given B
$P(A, B; C)$	Probability of A and B, given C
$P(A; B, C)$	Probability of A, given B and C
$\mathbf{F}$	Structure factor (vector)
F	Structure-factor amplitude
σ^2 or Σ	Variance of a Gaussian
$x \propto y$	x is proportional to y
$x \Rightarrow y$	x implies y

(*a*) (*b*) (*c*) (*d*)

Figure 1
Photographs of (*a*) four-sided die, (*b*) six-sided die, (*c*) ten-sided die and (*d*) eight-sided die

2.3. Dice, independence and log-likelihood

A game of dice with more data.

I put four unbiased dice in a box: one four-sided, one six-sided, one eight-sided and one ten-sided.

I select a die at random.

I roll that die three times and tell you the results of the rolls.

Which die did I most likely select?

If I roll a 7 three times, you would expect that I selected an eight-sided die, as the answer should be consistent with the game above when only one roll (of a 7) was made. How is the formal analysis performed? The chance of rolling a 7 three times from the four- or six-sided die is 0, but what is the chance of throwing a 7 three times from an eight-sided or ten-sided die? The chance of throwing a 7, or any other number, the second or third time is not influenced by the value of the first roll. This is the principle of independence. When probabilities are independent, they multiply. If the calculations are performed, the eight-sided die is indeed more likely,

$$P(7, 7, 7; \mathbf{8}) = \tfrac{1}{8} \times \tfrac{1}{8} \times \tfrac{1}{8} = \tfrac{1}{512},$$
$$P(7, 7, 7; \mathbf{10}) = \tfrac{1}{10} \times \tfrac{1}{10} \times \tfrac{1}{10} = \tfrac{1}{1000}.$$

After obtaining data from three rolls, your confidence that you have guessed the correct die has increased compared with when you only knew the result of one roll, so the likelihood ratio increases,

$$\frac{P(7, 7, 7; \mathbf{8})}{P(7, 7, 7; \mathbf{10})} = \frac{\frac{1}{512}}{\frac{1}{1000}} = 1.953.$$

What is the probability of rolling a 7 from an eight-sided die one hundred thousand times? (Of course, if you really were to roll 7 one hundred thousand times, you might have some difficulty believing that the die is unbiased. Please continue to assume that it is.) Although the formula for the probability can be written down,

$$P(7 \ldots \text{one hundred thousand times}; \mathbf{8}) = \frac{1}{8^{100\,000}},$$

and you could work out the answer and write it down on a (very long) piece of paper,

$$P(7 \ldots \text{one hundred thousand times}; \mathbf{8}) = 0.(\ldots 90308 \text{ zeroes} \ldots)10029997 \ldots,$$

the number is too small (has too many decimal places) to be stored by a computer. The solution to this computational problem is to calculate log-likelihood rather than the likelihood,

$$\log[P(7 \ldots \text{one hundred thousand times}; \mathbf{8})] = -90\,309.$$

Calculation of the log-likelihood solves the small-number computation problem, but is the switch from using the likelihood allowed? Fortunately it is, because logarithmic functions are monotonic functions [*i.e.* if $a < b$ then $\log(a) < \log(b)$]. This means that the parameter values obtained by optimizing log-likelihood are the same as the parameter values obtained by optimizing the likelihood. In fact, computer algorithms are designed to minimize, so parameters are optimized by minimizing the −log-likelihood. There are also other more theoretical justifications for using the log-likelihood, which come from the statistical field of information theory.

Is there a paradox in that the computer needs to store the likelihood before taking its logarithm? Fortunately not, because there is a shortcut to the log-likelihood when the total likelihood is a product of likelihoods, (*i.e.* when the likelihoods are independent),

$$\log\left(\prod_{i=1}^{N} P_i\right) = \sum_{i=1}^{N} \log(P_i).$$

In the case where I rolled 7 three times from an eight-sided die, there are thus two ways of calculating the log-likelihood. Using the product method,

$$\begin{aligned}\log[P(7, 7, 7; \mathbf{8})] &= \log\left[\prod_{i=1}^{3} P(7; \mathbf{8})_i\right] \\ &= \log(\tfrac{1}{8} \times \tfrac{1}{8} \times \tfrac{1}{8}) \\ &= \log(0.001953) \\ &= -2.70927.\end{aligned}$$

Using the sum method,

$$\begin{aligned}\log[P(7, 7, 7; \mathbf{8})] &= \sum_{i=1}^{3} \log[P(7; \mathbf{8})_i] \\ &= \log(\tfrac{1}{8}) + \log(\tfrac{1}{8}) + \log(\tfrac{1}{8}) \\ &= -0.90309 - 0.90309 - 0.90309 \\ &= -2.70927.\end{aligned}$$

However, the product method required the intermediate of calculating a number close to zero (0.001953), while the sum method did not require any numbers close to zero (the smallest numbers were the independent probabilities themselves, 0.125).

When the log-likelihood is used instead of the likelihood, the log-likelihood gain is calculated instead of the likelihood ratio. The log-likelihood gain is the difference between log-likelihoods [since $\log(a/b) = \log(a) - \log(b)$]. Whereas for the likelihood ratio more favourable likelihoods are indicated by values greater than 1, for the log-likelihood gain they are indicated by any positive value. The log-likelihood gain for the die being the eight-sided rather than the ten-sided after rolling 7 three times is

$$\begin{aligned}\log[P(7, 7, 7; \mathbf{8})] - \log[P(7, 7, 7; \mathbf{10})] &= (-2.70927) - (-3) \\ &= 0.29073.\end{aligned}$$

What would happen if the result of previous rolls influenced the result of the subsequent rolls? In this case the data points are not independent, but correlated. Note that correlation is not the same as bias. A biased die would be one that, for example, always rolled a 7, but a correlated die would be one that, for example, always rolled one number higher than the previous roll. Highly correlated data points make the determination of the likelihood difficult, if not impossible, and so the assumption of independence is often applied even when it is not justified. In crystallography, reflections are assumed to

be independent, even though to a certain extent they are not. Correlations are introduced by the presence of solvent, which means that the molecular transform is over-sampled, and by non-crystallographic symmetry (if present). However, the correlations are sufficiently weak that the approximation of assuming independence is very good. To calculate the total log-likelihood for all the reflections in a data set (of the order of one hundred thousand), the sum of the log-likelihoods for each reflection is used.

2.4. Dice and Bayes' theorem

A game of dice with multiple copies of a die.
I put one eight-sided die and eight ten-sided dice in a box.
I select a die at random.
I roll the die and tell you the result of the roll.
Which die did I most likely select?

I roll a 4. In this case the probability of selecting the ten-sided die in the first place overwhelms the slightly higher chance of rolling the 4 from the eight-sided die. The chance of selecting the ten-sided die in the first place is included in the probability calculation with 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, so Bayes' theorem becomes

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

Bayes' theorem is also called the rule of inverse probability since it shows how to turn P(data; model) (*e.g.* the probability of rolling the 4 from the ten-sided die, which we can calculate) into P(model; data) (*e.g.* the probability of the ten-sided die given a roll of 4, which is what we want to know). P(model) is the probability of the model without having any data (*e.g.* the chance of selecting the ten-sided die in the first place). P(model; data) is called the posterior probability, P(data; model) is called the likelihood (as before) and P(model) is called the prior probability. If Bayes' theorem is used to calculate the probability rather that just the likelihood, then the method of optimizing the probability should properly be called the maximum-posterior method, rather than the maximum-likelihood method, but the term 'maximum likelihood' is generally used for both. True maximum likelihood can be thought of as a special case of maximum posterior when the prior probability P(model) is constant for all the models. This was the case for the examples in §§2.2 and 2.3 above.

Using Bayes' theorem, the probability that the die was ten-sided given a roll of 4,

$$\begin{aligned} P(\mathbf{10};4) &= P(\mathbf{10}) \times P(4;\mathbf{10}) \\ &= \tfrac{8}{9} \times \tfrac{1}{10} \\ &= 0.0\bar{8}, \end{aligned}$$

is higher than the probability that the die was eight-sided given a roll of 4,

$$\begin{aligned} P(\mathbf{8};4) &= P(\mathbf{8}) \times P(4;\mathbf{8}) \\ &= \tfrac{1}{9} \times \tfrac{1}{8} \\ &= 0.013\bar{8}, \end{aligned}$$

so a ten-sided die is more likely, as expected.

Bayes' theorem is very useful in crystallography because it enables exploitation of the things that are known about protein structure even before the X-ray data are collected. For example, a carbon–oxygen double bond is known to be 1.23 Å long. So, if the electron density for a structure showed no density 1.23 Å from a particular peptide carbon, but a large piece of density 2 Å away from it, prior knowledge of the carbon–oxygen double-bond length means that it would be extremely unlikely that the density 2 Å away was due to an O atom bound to the carbonyl O atom. It would be more likely that the density 2 Å away from the carbon was due to noise or some other feature of the structure. However, if the O atom had been moved into this density during rebuilding (and the carbon–oxygen bond stretched), a refinement program would use Bayes' theorem to restrain the bond length to 1.23 Å and produce the more likely structure. Bayes' theorem is also used in density modification, where information about solvent content, non-crystallographic symmetry *etc.* is introduced (Terwilliger, 2000; McCoy, 2002).

2.5. Dice and integrating out nuisance variables

A game of dice with unknown dice.
I put a six-sided and an eight-sided die in a red box and a four-sided and ten-sided die in a blue box.
I select a die from each of the red and blue boxes at random and put them in a yellow box.
I select a die at random from the yellow box, roll the die and tell you the result.
Did the die most likely come from the red box or the blue box originally?

I roll a 3. The problem here is to calculate the likelihoods P(3; blue box) and P(3; red box) and find the maximum without knowing which die actually produced the roll of 3.

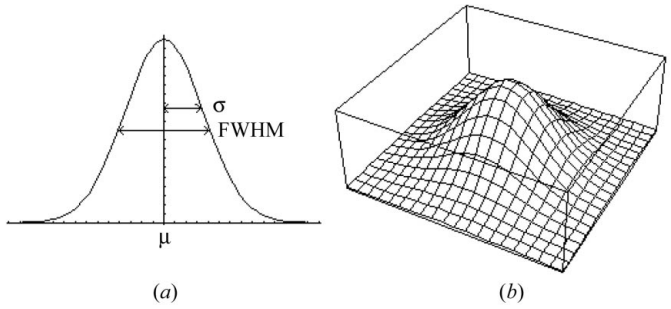

Figure 2
(*a*) One-dimensional Gaussian $\{1/[(2\pi)^{1/2}\sigma]\}\exp[-(x-\mu)^2/2\sigma^2]$ and (*b*) radially symmetric two-dimensional Gaussian $[1/(2\pi\sigma^2)]\exp[-(\mathbf{x}-\boldsymbol{\mu}|^2)/2\sigma^2]$. The mean is μ (one-dimensional)/$\boldsymbol{\mu}$ (two-dimensional). The standard deviation is σ, which is half-width at ∼61% of the peak height. The variance is σ^2. The full-width half-maximum (FWHM) = 2.35σ. The area (one-dimensional)/volume (two-dimensional) under the curve is 1.

Consider $P(3;$ blue box$)$. The blue box could have contained either the four-sided or the ten-sided die. To calculate $P(3;$ blue box$)$, the likelihood of the 3 being rolled from the two possibilities for the contents of the blue box (the four-sided and the ten-sided die) are added,

$$P(3;\text{ blue box}) = P(3, \mathbf{4};\text{ blue box}) + P(3, \mathbf{10};\text{ blue box})$$

The basic probability identity $P(A, B) = P(B; A) \times P(A)$ [which can also have any number of conditions added, so $P(A, B; C) = P(B; A, C) \times P(A; C)$, for example] can be used to write

$$\begin{aligned} P(3;\text{ blue box}) = {} & P(3; \mathbf{4},\text{ blue box}) \times P(\mathbf{4};\text{ blue box}) \\ & + P(3; \mathbf{10},\text{ blue box}) \times P(\mathbf{10};\text{ blue box}). \end{aligned}$$

Substituting in values for these probabilities,

$$P(3;\text{ blue box}) = (\tfrac{1}{4} \times \tfrac{1}{2})(\tfrac{1}{10} \times \tfrac{1}{2}) = 0.175.$$

Likewise, $P(3;$ red box$)$ is the likelihood of the 3 being rolled and the die being six-sided plus the likelihood of the 3 being rolled and the die being eight-sided,

$$\begin{aligned} P(3;\text{ red box}) &= P(3, \mathbf{6};\text{ red box}) + P(3, \mathbf{8};\text{ red box}) \\ &= P(3; \mathbf{6},\text{ red box}) \times P(\mathbf{6};\text{ red box}) \\ &\quad + P(3; \mathbf{8},\text{ red box}) \times P(\mathbf{8};\text{ red box}) \\ &= (\tfrac{1}{6} \times \tfrac{1}{2}) + (\tfrac{1}{8} \times \tfrac{1}{2}) \\ &= 0.14583\bar{} \end{aligned}$$

So, it is slightly more likely that the die came from the blue box if I roll a 3.

The likelihood for the red and blue boxes has been calculated even though which die actually produced the roll of 3 was not known. Summing the probabilities for all the possibilities for the die solves the problem of not knowing which die actually produced the roll. In general, if the unknown variable (call it x) of the model can take n values between a and b, then

$$P(\text{data};\text{model}) = \sum_{i=1}^{n} P(\text{data}, x_i;\text{ model}),$$

where $a < x_i \leq b$.

The probability distribution for the dice is for discrete variables, because it is only defined for certain values (the dice must have an integer number of sides). In crystallography, the probability distributions are for continuous variables, meaning that they are defined for all values (an infinite number) over an interval (for example, an atom can be anywhere and an occupancy can be anywhere between 0 and 1). When the probability distribution is continuous, the sum in the equation for the discrete probability distribution becomes an integral, because an integral can be thought of as a sum of an infinite number of infinitesimally small numbers. If the unknown variable x can take values between a and b, then

$$P(\text{data};\text{model}) = \int_{a}^{b} P(\text{data}, x;\text{ model})\ \mathrm{d}x.$$

The unknown variable x is called a nuisance variable. The removal of a nuisance variable from a probability distribution by integration is called integrate out (or marginalization of) the nuisance variable. Although termed 'nuisance', these variables can be very useful in probability distributions. It may be easier to describe a probability function using an extra variable (such as the phase of the observed structure factor) and then to integrating it out at the end of the analysis than to attempt to develop a probability function without ever referring to the extra variable.

2.6. Dice and the central limit theorem

A game of dice taking the average of many rolls of the dice.

I have a six-sided die.

I roll the die 40 times and add up the values of the rolls, then divide the sum by 40.

I do this 10 000 times, plotting the final average value from each game on a histogram.

What form does the histogram take?

The histogram is Gaussian (bell-shaped curve, see Fig. 2), with a maximum at 3.5 (see Fig. 3). Now I play the game again with a biased six-sided die: the die is biased so that it will roll its values with a probability linearly proportional to the value, *i.e.* a 2 is twice as likely as a 1, a 3 is three times as likely as a 1 *etc*. Again, the histogram looks like a Gaussian. The only difference is that the mean of the distribution is shifted to about 4.3 and the variance of the distribution is smaller (see Fig. 2 for an explanation of the mean and variance of a Gaussian). Now I play the game again with a six-sided die that is biased so that it will roll its values with a probability proportional to the square of the value *i.e.* a 2 is four times more likely than a 1, a 3 is nine times more likely than a 1 *etc*. Again, the histogram looks like a Gaussian (with an even higher mean and smaller variance). For most types of bias of the die, the histogram generated by the game of dice is Gaussian, even when the bias of the die (from which the average is computed) is decidedly non-Gaussian. This property is called the central limit theorem. The central limit theorem is possibly the most important theorem in probability. In crystallography, the central limit theorem allows us to describe the errors in the structure factors (in reciprocal space) that arise from errors in the atomic model (in real space). It says that even though the errors in an individual atom's contribution to the total structure factor may be very complicated, in the end the error for the total structure factor (the sum of the atomic structure-factor contributions) is a simple two-dimensional Gaussian in reciprocal space.

2.7. Dice summary

Maximum likelihood: the best model is the one that maximizes the probability of observing the experimental data.

Independence: probabilities multiply when the experimental data points are independent, *i.e.* all observations are independent of one another.

Log-likelihood: the log-likelihood is used instead of the likelihood as it has its maximum at the same parameter values

as the likelihood but it is safer to calculate on a computer because the numerical range is smaller.

Bayes' theorem: P(model; data) = P(model) × P(data; model), where P(data; model) is called the likelihood and P(model) is called the prior probability.

Integrating out variables: nuisance variables in a joint probability distribution can be eliminated by integration.

Central limit theorem: the distribution of the average tends to be Gaussian, even when the distribution from which the average is computed is decidedly non-Gaussian.

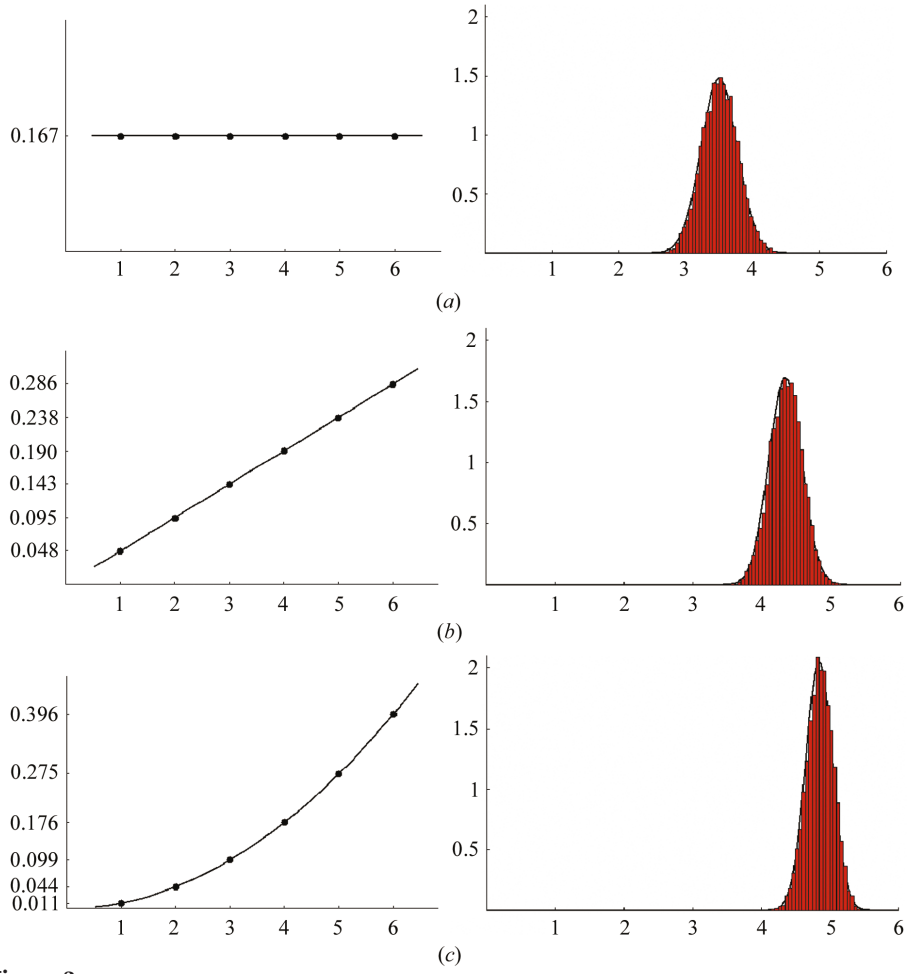

Figure 3
Central limit theorem. On the left are the probability distributions showing the bias of the six-sided die (shown with dots) and on the right the histogram of 10 000 trials of the average of 40 rolls of the die. If the probability distribution is continuous rather than discrete (shown on the right with a line connecting the dots), the distribution of the average is also continuous (shown with Gaussian function over the histogram). (*a*) Unbiased die: the distribution of the average is a Gaussian with $\mu = 3.5$ and $\sigma = 0.27$. (*b*) Linearly biased die: the distribution of the average is a Gaussian with $\mu = 4.3$ and $\sigma = 0.24$. (*c*) Quadratically biased die: the distribution of the average is a Gaussian with $\mu = 4.8$ and $\sigma = 0.19$.

3. Maximum likelihood in macromolecular crystallography

There are many ways of applying maximum likelihood to crystallography. Ideally, all the information from chemistry and the diffraction experiment should be included to create the 'mother of all likelihood functions'. Although the chemical and diffraction information that should contribute to this likelihood function is known, there are too many correlations between the contributions to allow a practical precise formula to be written down. This is rather unfortunate, because there is enough information in the chemistry and the diffraction experiment to solve the phase problem *ab initio* (*cf.* direct methods; Bricogne, 1993). Instead, simplifications and approximations are made to allow maximum likelihood to be applied to specific areas of crystallography such as refinement, molecular replacement and experimental phasing.

4. Refinement

The Bayesian view of crystallographic refinement is that the prior probability comes from chemistry (a great deal is known about what molecules look like even before the experiment) and the likelihood comes from the X-ray diffraction experiment (Pannu & Read, 1996; Bricogne & Irwin, 1996; Murshudov *et al.*, 1997). The probability function for refinement (here called P-refinement) is thus, by Bayes' theorem (see §2.4), the product of the prior probability (here called P-chemistry) and the likelihood (here called P-Xray),

$$P(\text{model; data}) = P(\text{model}) \times P(\text{data; model})$$
$$\Rightarrow P\text{-refinement} = P\text{-chemistry} \times P\text{-Xray}.$$

The chemical probabilities for all the different chemical interactions in the structure are taken to be independent (see §2.3), so that P-chemistry is the product of these individual chemical interaction probabilities P-chemistry$_i$. This is not a very good approximation, as the bond lengths and angles are correlated with each other; the problems that this approximation causes are discussed in §4.4. If the number of interactions is I,

$$P\text{-chemistry} = \prod_{i=1}^{I} P\text{-chemistry}_i.$$

It is also assumed that reflections are independent, so that P-Xray is the product of the reflection likelihoods (P-Xray$_r$).

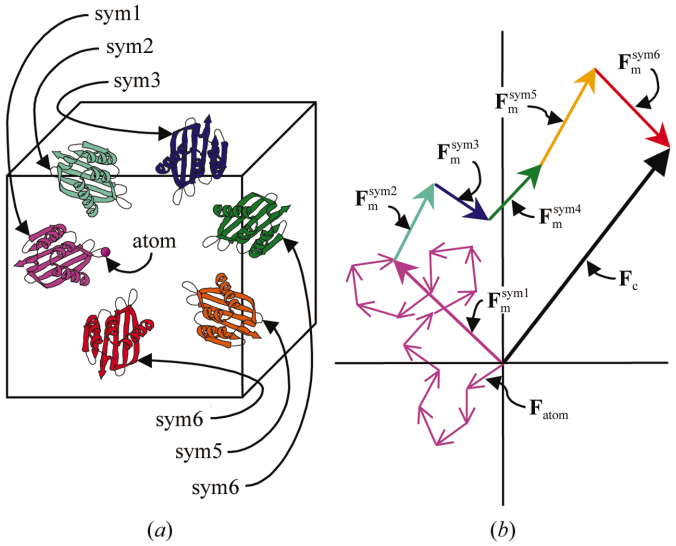

Figure 4
Total structure factor for all scattering in unit cell. (*a*) The unit cell contains six symmetry-related molecules. (*b*) The total structure factor ($\mathbf{F}_c$) for a reflection is built up by adding the structure-factor contributions from the atoms in a molecule ($\mathbf{F}_{atom}$) to give the structure factor for the molecules ($\mathbf{F}_m$) and then adding all the symmetry-related structure-factor contributions for all the molecules in the unit cell.

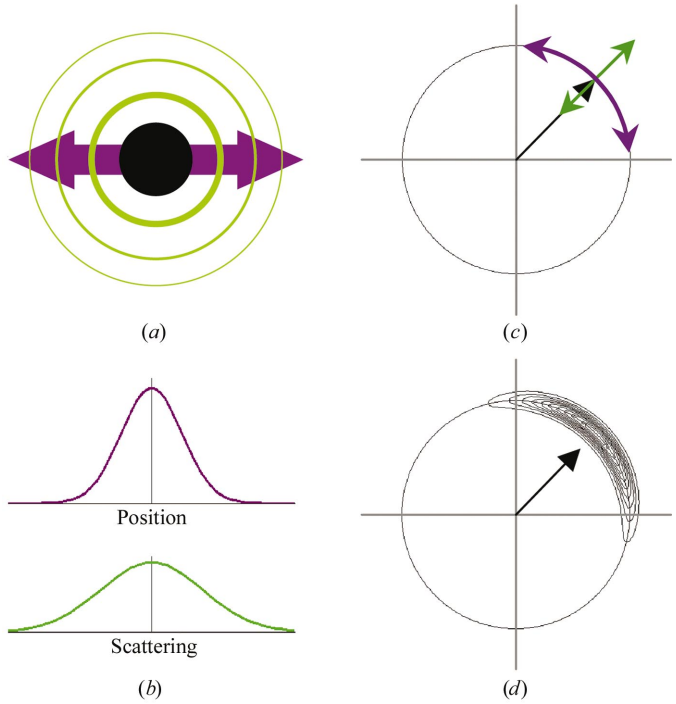

Figure 5
Errors for an atomic structure factor. (*a*) An atom has variation in position (indicated by purple arrow) and in scattering (indicated in green concentric circles). (*b*) The variation in the atom's position and scattering are Gaussian. (*c*) The atom at its mean position with its mean scattering has a structure factor $\mathbf{F}_{atom}$ (shown with a black vector). Variation in the atom's position corresponds to variation in the phase of $\mathbf{F}_{atom}$ (shown with a purple arrow) and variation in the scattering corresponds to variation in the length of $\mathbf{F}_{atom}$ (shown with a green arrow). (*d*) The distributions of the structure factors owing to variation in the atom's position and scattering combine to give a boomerang-shaped structure-factor distribution (indicated with black contours). Since the distribution of structure factors is symmetric about $\mathbf{F}_{atom}$, the average structure factor is shorter than $\mathbf{F}_{atom}$ (by a fraction d, where $0 < d < 1$) but in the same direction as $\mathbf{F}_{atom}$ ($d\mathbf{F}_{atom}$).

This is a good approximation (see §2.3). If the number of reflections is R,

$$P\text{-Xray} = \prod_{r=1}^{R} P\text{-Xray}_r.$$

Since there are hundreds of thousands of interactions and hundreds of thousands of reflections, the log-likelihood is calculated rather than the likelihood (see §2.3). To optimize the model parameters (atomic positions, occupancies and B factors), the −log-likelihood is minimized,

$$-\log P\text{-refinement} = -\sum_{i=1}^{I} \log P\text{-chemistry}_i - \sum_{r=1}^{R} \log P\text{-Xray}_r.$$

4.1. Prior probability

Macromolecules obey the same chemical rules as small organic molecules and so ideal bond lengths and angles for macromolecules can be derived from the results of small-molecule crystallography (Engh & Huber, 1991). The bond lengths and angles in the crystal are restrained to these ideal values using a probability distribution. For example, the prior probability for having a bond of length b is given by a Gaussian about the ideal length b_{ideal} for the bond type (see Fig. 2 for the equation for a Gaussian),

$$P\text{-chemistry}_{bond} = \frac{1}{(2\pi)^{1/2}\sigma_b} \exp\left[-\frac{(b - b_{ideal})^2}{2\sigma_b^2}\right].$$

Here, σ_b reflects the distribution of a particular bond type about its mean; *e.g.* a C—C bond has an ideal length of 1.54 Å

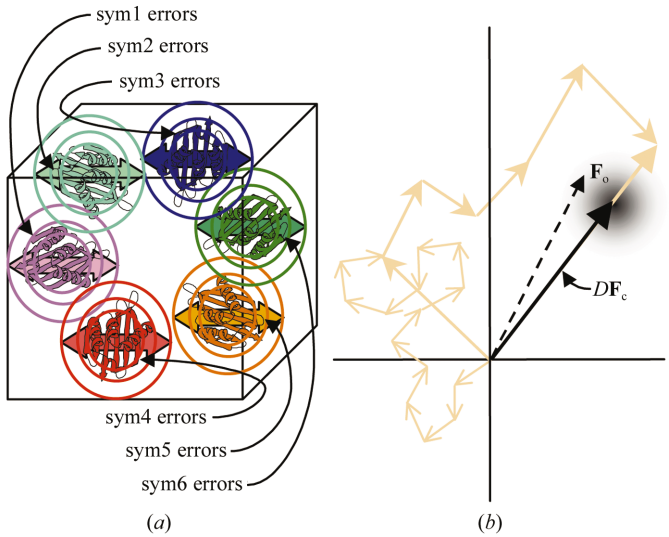

Figure 6
Errors in the total structure factor. (*a*) The unit cell contains six symmetry-related molecules. The atoms have errors in their positions and scattering, indicated by the arrows and concentric circles, respectively. (*b*) By the central limit, the probability distribution for the sum of the structure-factor contributions from all the atoms is a two-dimensional Gaussian in reciprocal space, centred at $D\mathbf{F}_c$, where $0 < D < 1$, shown with grey shading. The structure-factor contributions from atoms and molecules as in Fig. 4 are shown in pink.

and a σ_b of about 0.02 Å. There are similar equations for the other types of chemical interaction restraints.

4.2. Refinement likelihood

The likelihood for a reflection (P-X-ray$_r$) is the probability of the data (*i.e.* the observed structure-factor amplitude F_o) given the current model. The model is in real space and the X-ray observed data are in reciprocal space, so in order to calculate the likelihood, the model (in real space) must be used to calculate structure factors (in reciprocal space). The structure factor for the whole unit cell ($\mathbf{F}_c$) is calculated as follows: first the structure factor for the model in the asymmetric unit ($\mathbf{F}_m$) is calculated from the sum of the structure factors of the atoms in that model ($\mathbf{F}_{atom}$). Then, $\mathbf{F}_m$ and all its symmetry relatives are added to obtain the total structure factor $\mathbf{F}_c$ (see Fig. 4; the importance of the symmetry relatives will become apparent in the explanation of the rotation-function likelihood below). However, the data for a given reflection is the observed structure-factor amplitude F_o, so in order to compare like with like the model must be the calculated structure-factor amplitude F_c,

$$P\text{-Xray}_r = P(\text{data}; \text{model}) = P(F_o; F_c).$$

Without considering errors, if F_c matches F_o the probability is 1 and if it does not match the probability is 0 (the model is either 'correct' or 'incorrect'). However, if errors in the model and the data are considered, then F_c and F_o are allowed to differ somewhat and the likelihood function should give a non-zero probability when F_c and F_o are close (the closer the better). It is easier to model the errors in terms of the phased structure factors $\mathbf{F}_c$ and $\mathbf{F}_o$ with the phase between them defined as α, rather than in terms of the structure-factor amplitudes alone. The introduced variable, the phase α, is a nuisance variable (a case where a nuisance variable is very useful) and must be integrated out of the probability distribution at the end of the analysis (see §2.5). The integration limits are 0–2π, *i.e.* all angles,

$$P\text{-Xray}_r = \int_0^{2\pi} P(F_o, \alpha; F_c)\, d\alpha.$$

The errors in the model arise from Gaussian errors in the atomic positions and atomic scattering. Gaussian errors in the atomic positions and scattering give rise to Gaussian errors in the phases and amplitudes of the corresponding atomic structure-factor contributions, respectively (see Fig. 5). When these atomic structure-factor contributions and their errors are summed to give the total structure factor and its error for a given reflection, by the central limit theorem (see §2.6) the resulting distribution is a two-dimensional Gaussian (see Fig. 2) in reciprocal space centred on $D\mathbf{F}_c$ (where D is between 0 and 1) with variance termed σ_Δ^2 (see Fig. 6),

$$P(\mathbf{F}_o; \mathbf{F}_c) = \frac{1}{\pi\sigma_\Delta^2} \exp\left(-\frac{|\mathbf{F}_o - D\mathbf{F}_c|^2}{\sigma_\Delta^2}\right).$$

Using this probability and the integral above, it can be shown (Appendix A) that the likelihood function is a Rice distribution (Sim, 1959; Read, 1990),

$$P\text{-Xray}_r = \frac{2F_o}{\sigma_\Delta^2} \exp\left(-\frac{F_o^2 + D^2F_c^2}{\sigma_\Delta^2}\right) I_0\left(\frac{2F_oDF_c}{\sigma_\Delta^2}\right),$$

where I_0 is the modified Bessel function of order 0. The Rice distribution is the key distribution for maximum likelihood in crystallography and it will appear over and over again in the equations below. It applies to acentric reflections (those for which the phase is not restricted) and, for simplicity, the discussions below will only concern acentric structure factors (and assume the expected intensity factor, generally denoted ε, to be 1). For a full explanation of the derivation of the Rice function, see Appendix A. Centric structure factors (those where the phase is restricted to 0 or 180°) are treated similarly to give a different likelihood function (see Appendix B).

There are also experimental errors (σ_F) in the measurements. Experimental error is accounted for by widening the probability distribution, a method that is termed inflating the variance (Green, 1979; Murshudov *et al.*, 1997; de La Fortelle & Bricogne, 1997). The likelihood function used for refinement is therefore given by

$$P\text{-Xray}_r = \frac{2F_o}{\sigma_\Delta^2 + \sigma_F^2} \exp\left(-\frac{F_O^2 + D^2F_C^2}{\sigma_\Delta^2 + \sigma_F^2}\right) I_0\left(\frac{2F_ODF_C}{\sigma_\Delta^2 + \sigma_F^2}\right). \quad (1)$$

4.3. Sigma A

D and σ_Δ are anticorrelated: if the model is very bad and therefore if σ_Δ^2 is large then D will be small and *vice versa*. If E values (normalized structure factors) are used rather than F values, D and σ_Δ can be replaced with a single parameter σ_A (Srinivasan & Ramachandran, 1965), with $DF_c = \sigma_A E_c$ and $\sigma_\Delta^2 = 1 - \sigma_A^2$,

$$\begin{aligned} P\text{-Xray}_r &= P(E_o; E_c) \\ &= \frac{2E_o}{1 - \sigma_A^2 + \sigma_E^2} \exp\left(-\frac{E_o^2 + \sigma_A^2E_c^2}{1 - \sigma_A^2 + \sigma_E^2}\right) I_0\left(\frac{2E_o\sigma_AE_c}{1 - \sigma_A^2 + \sigma_E^2}\right), \end{aligned}$$

where σ_E is the normalized structure-factor experimental error. The probability distributions are very sensitive to the estimates of σ_A, and σ_A is refined along with the atomic parameters in structure refinement. Unfortunately, if the same data are used to refine σ_A and the atomic parameters, the data are severely overfitted and σ_A is overestimated. This problem is partially avoided by estimating σ_A from the data that are used to compute R_{free} (which are excluded from the refinement).

4.4. Weighting

In principle, if all errors are estimated properly there is no need to apply a weighting between the prior probability (P-chemistry) and likelihood (P-Xray) to calculate P-refinement using Bayes' theorem, but in practice it is necessary to overweight the likelihood (P-Xray) for refinement to converge. This is partly because the probabilities used are only

approximate (particularly for the chemistry terms, where the correlations between bond lengths and angles are not taken into account) and partly because the refinement algorithm does not account for the fact that improvements in the model will sharpen the experimental likelihood function (because the model and the σ_A values are refined against different subsets of the data). As the resolution becomes higher and the model becomes better, the amount of over-weighting required is reduced.

4.5. Experimental phases

Experimental phasing information can be incorporated into the refinement likelihood function as a prior probability when integrating out the phase (Pannu *et al.*, 1998). The prior probability can be modelled using Hendrickson–Lattman coefficients (Hendrickson & Lattman, 1970).

4.6. Probabilities and energies

Some refinement programs minimize energy rather than the −log-likelihood. In fact, the two targets of refinement are equivalent. If the experiment is considered as a physical system with energy, Boltzmann's law gives the probability P of observing a state in the physical system with energy E,

$$P \propto \exp(-E/kT),$$

where k is Boltzmann's constant and T is the temperature. Taking the logarithm of Boltzman's law, the energy is proportional to the logarithm of the probability,

$$E \propto -kT \ln P.$$

Boltzman's law in logarithm form leads to harmonic bond restraints,

$$\begin{aligned} P\text{-chemistry}_{\text{bond}} &\propto \exp\left[-\frac{(b-b_{\text{ideal}})^2}{2\sigma_b^2}\right] \\ \Rightarrow E\text{-chemistry}_{\text{bond}} &\propto -kT\frac{(b-b_{\text{ideal}})^2}{2\sigma_b^2}. \end{aligned}$$

Boltzman's law in this logarithm form also allows Bayes' theorem (in terms of probabilities) to be expressed in terms of energies

$$\begin{aligned} &P\text{-refinement} = P\text{-chemistry} \times P\text{-Xray} \\ &\Rightarrow E\text{-refinement} = E\text{-chemistry} + E\text{-Xray}. \end{aligned}$$

5. Molecular replacement

Maximum-likelihood molecular replacement (Bricogne, 1992; Read, 2001, 2003*b*) can be divided into a rotation function (RF) followed by a translation function (TF) in the same way that traditional molecular-replacement methods are. Each type of search is a 'brute-force' search procedure. The likelihood for the models is generated on a grid of angles (RF) or positions (TF) and the angle (RF) or position (TF) of the model that has the highest likelihood is selected as the 'solution' to the molecular-replacement problem. Currently, prior information (such as packing constraints and non-crystallographic symmetry) is not included in maximum-likelihood molecular replacement and so Bayes' theorem (see §2.4) is not used. Reflections are assumed to be independent, so that the likelihood for the rotation function (here called P-RF) and the likelihood for the translation function (here called P-TF) is the product of the reflection likelihoods (see §2.3). If the number of reflections is R,

$$P\text{-RF} = \prod_{r=1}^{R} P\text{-RF}_r$$

and

$$P\text{-TF} = \prod_{r=1}^{R} P\text{-TF}_r.$$

In practice, the −log-likelihood is used as the target for the molecular-replacement searches,

$$-\log P\text{-RF} = -\sum_{r=1}^{R} \log P\text{-RF}_r$$

and

$$-\log P\text{-TF} = -\sum_{r=1}^{R} \log P\text{-TF}_r.$$

5.1. Translation-function likelihood

The data are the observed structure-factor amplitudes (F_{o}) and the model is the molecular-replacement structure oriented and positioned at the search point. This is exactly the same situation as for refinement: the approximate locations of all the atoms are known and a structure-factor amplitude F_{c} can be calculated from the scattering in the unit cell. The translation function target is therefore the same Rice function as the target for maximum-likelihood structure refinement. The only difference is that the errors will be much larger for the translation function than for refinement (D will be smaller and σ_Δ larger). The same function is also suitable for a brute-force six-dimensional (orientation and position) search,

$$\begin{aligned} P\text{-TF}_r &= P\text{-Xray}_r \\ &= \frac{2F_{\text{o}}}{\sigma_\Delta^2+\sigma_F^2}\exp\left(-\frac{F_{\text{o}}^2+D^2F_{\text{c}}^2}{\sigma_\Delta^2+\sigma_F^2}\right)I_0\left(\frac{2F_{\text{o}}DF_{\text{c}}}{\sigma_\Delta^2+\sigma_F^2}\right). \end{aligned}$$

5.2. Rotation-function likelihood

At each rotation-function search orientation, the model consists of the molecular-replacement model with defined orientation but undefined position. Undefined positions in real space correspond to undefined phases of the structure-factor contributions in reciprocal space. Thus, $\mathbf{F}_{\text{c}}$ cannot be calculated from the sum of the phased structure-factor contributions as it was for the case of refinement and the translation function. However, because the relative positions of the atoms in the model are known, the atomic structure-factor contributions ($\mathbf{F}_{\text{atom}}$) for the model can be added up

with relative phases to calculate F_m, *i.e.* the amplitude but not the phase of the model structure-factor contribution. This can also be performed for all the symmetry relatives of the model in order to obtain the set of amplitudes of the model structure-factor contributions, $\{F_m\}_{sym}$. The symmetry relatives have different amplitudes because as the model rotates its strength of scattering in any given direction changes. Since these model structure-factor contributions are unphased, they cannot be added together to obtain the structure factor for the scattering from the whole unit cell, $\mathbf{F}_c$. The model in reciprocal space for the rotation function is therefore not $\mathbf{F}_c$, but the set of amplitudes of the model structure-factor contributions, $\{F_m\}_{sym}$,

$$P\text{-RF}_r = P(\text{data; model}) = P(F_o; \{F_m\}_{sym}).$$

It is easiest to generate a function for this probability by introducing a (useful) nuisance phase variable, the phase α between the observed structure factor F_o and one of the F_m. It is best to select the symmetry relative of F_m with the largest amplitude, F_{big} (the reason is given shortly). The symmetry operator that gives rise to the largest F_m will be different for each reflection, so F_{big} corresponds to a different symmetry operator for each reflection. The set of symmetry relatives of F_m is thus split into the set not including F_{big}, $\{F_m\}_{sym\neq big}$, which are left unphased, and F_{big}, which is given the phase α relative to F_o. The introduced nuisance phase α must be integrated out of the probability distribution at the end of the analysis (see §2.6),

$$P\text{-RF}_r = \int_0^{2\pi} P(F_o, \alpha; \{F_m\}_{sym\neq big}, F_{big})\, d\alpha.$$

The probability distribution for $\{F_m\}_{sym\neq big}$ comes from a 'random walk' (Fig. 7) in reciprocal space. Fixing the phase of the largest of the symmetry relatives of F_m results in the narrowest probability distribution for the 'random walk' and this is why the largest F_m was chosen to have the phase α relative to F_o. Errors in the model must also be accounted for in the probability distribution, just as they were for refinement. Using the same reasoning that applied for developing the refinement target (see Fig. 6), errors in the model mean that all symmetry relatives of F_m (including F_{big}) are down-weighted by a D-factor ($0 \leq D \leq 1$). The probability distribution is thus a two-dimensional Gaussian centred on $D\mathbf{F}_{big}$ with variance Σ_S dependent on $\{DF_m\}_{sym\neq big}$ (see Fig. 8),

$$P(\mathbf{F}_o; \{F_m\}_{sym\neq big}, \mathbf{F}_{big}) = \frac{1}{\pi\Sigma_S}\exp\left(-\frac{|\mathbf{F}_o - D\mathbf{F}_{big}|^2}{\Sigma_S}\right).$$

Using this probability and the integral above, it can be shown (Appendix *A*) that the likelihood function is another Rice distribution,

$$P\text{-RF}_r = \frac{2F_o}{\Sigma_S}\exp\left(-\frac{F_o^2 + D^2F_{big}^2}{\Sigma_S}\right)I_0\left(\frac{2F_oDF_{big}}{\Sigma_S}\right).$$

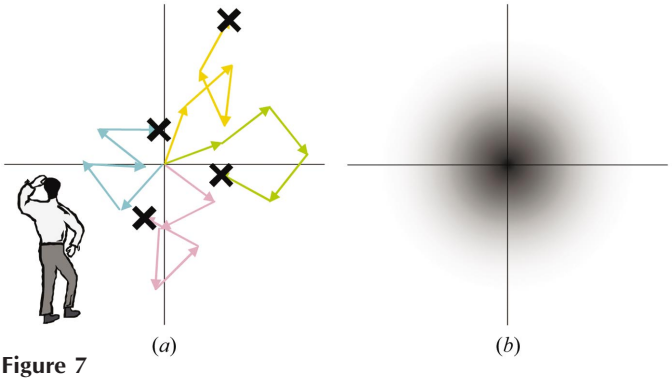

Figure 7
Random walk. (*a*) Starting at the origin, a walker takes N (in this case, five) steps, each step in a random direction. Each step is the same length and independent of the previous one. Because of the many random choices, the final position of the walker varies each time. Four final positions are shown (marked ×). Some final positions are more likely than others. (*b*) The probability that the walker will be at a given final position after N steps is proportional to a two-dimensional Gaussian, shown with grey shading.

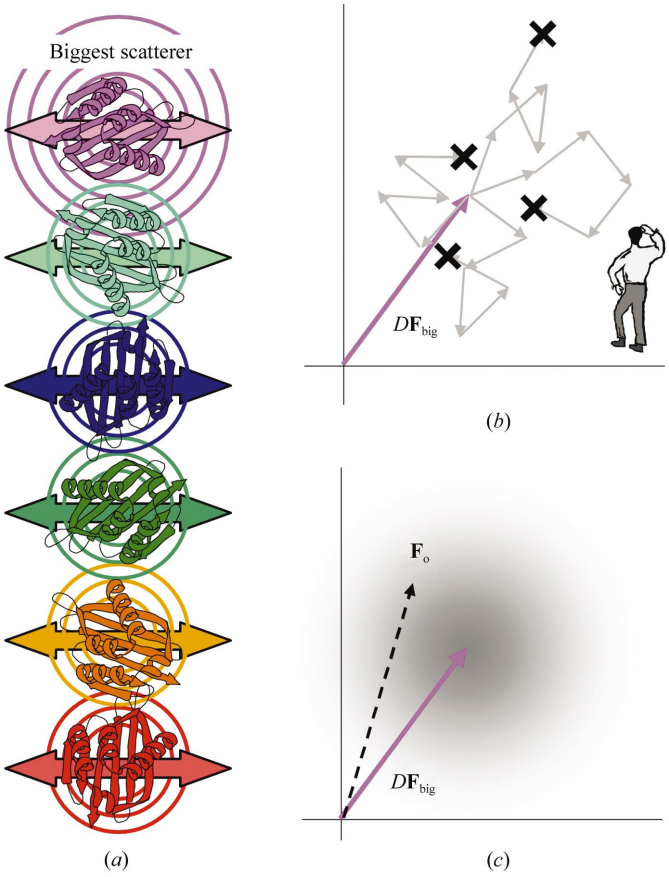

Figure 8
Rotation-function likelihood. (*a*) The unit cell contains six symmetry-related molecules. For a given orientation of the search, the orientation but not the position of the six molecules is defined. Therefore, the amplitudes but not the phases of the six corresponding structure-factor contributions are defined. The atoms in the molecules have errors in their positions and scattering, indicated by the arrows and concentric circles, respectively. The molecule in the orientation giving the largest scattering is shown in magenta. (*b*) The largest model structure-factor contribution, F_{big}, is given a phase α relative to the observed structure factor F_o, with the other model structure-factor contributions making a 'random walk' around the end of this one phased structure-factor contribution. (*c*) The resulting probability distribution for $\mathbf{F}_o$ is a two-dimensional Gaussian centred on $D\mathbf{F}_{big}$, shown with grey shading.

Experimental errors (σ_F) are incorporated by inflating the variance of the distribution, as was the case for the refinement likelihood function,

$$P\text{-RF}_r = \frac{2F_o}{\Sigma_S + \sigma_F^2}\exp\left(-\frac{F_o^2 + D^2F_{big}^2}{\Sigma_S + \sigma_F^2}\right)I_0\left(\frac{2F_O DF_{big}}{\Sigma_S + \sigma_F^2}\right). \quad (2)$$

Notice the similarities between this equation and the equation for P-Xray$_r$ [and P-TF$_r$; (1)]. The only differences are that F_{big} replaces F_c and Σ_S replaces σ_Δ^2. The latter difference shows an unfortunate inconsistency in the notation for variances that has arisen in crystallography. Sometimes the variance is shown as the square of the standard deviation, with the standard deviation written with a lower case Greek sigma (*e.g.* σ_Δ^2), and sometimes the variance is shown as a single parameter, written with an upper case Greek sigma (*e.g.* Σ_S). The differences in the equations can be traced back to differences in the position of the centre and difference in the width of the two-dimensional Gaussian in reciprocal space that gave rise to the Rice distribution.

5.3. Partial structure

Maximum-likelihood molecular replacement allows incorporation of any information about the structure already determined, *e.g.* known orientation and position of partial structure, known orientation of partial structure only and any combination thereof. Any partial structure information makes the probability distribution more exacting (reduces the variance) and improves the signal of the search.

5.4. Fast searches

Maximum-likelihood brute-force rotation and translation searches are very slow to compute. However, there are approximations to the full search targets that can be calculated with fast Fourier transforms and are therefore much faster. The fast rotation search is calculated with a series of two-dimensional fast Fourier transforms, while the fast translation search is calculated with one three-dimensional fast Fourier transform. These likelihood-enhanced fast rotation and translation searches can be generated by a Taylor series expansion of the full likelihood targets (Storoni *et al.*, 2004).

6. Experimental phasing

There are many forms of experimental phasing, including MIR (multiple-wavelength isomorphous replacement), MIRAS (multiple-wavelength isomorphous replacement with anomalous scattering), MAD (multiple-wavelength anomalous dispersion) and SAD (single-wavelength anomalous dispersion). They all have different types of data and types of models and so require different types of likelihood functions (Bricogne, 1991; Read, 1991, 1994; de La Fortelle & Bricogne, 1997). Prior information is not included in maximum-likelihood experimental phasing and so Bayes' theorem is not used (see §2.4). Reflections are assumed to be independent, so that the total likelihood is the product of reflection likelihoods (see §2.3). If the number of reflections is R, then for example in the case of MIR the likelihood (here called P-MIR) is given by

$$P\text{-MIR} = \prod_{r=1}^{R} P\text{-MIR}_r.$$

In practice, the −log-likelihood is used,

$$-\log P\text{-MIR} = -\sum_{r=1}^{R} \log P\text{-MIR}_r.$$

Similarly, the MIRAS, MAD and SAD likelihoods are the products of their respective reflection likelihoods. The heavy-atom sites must have been found using a Patterson, direct-methods or dual-space method before invoking maximum-likelihood phasing. The heavy-atom sites (in real space) are then used to calculate the model for maximum likelihood, the heavy-atom structure factors $\mathbf{F}_H$ (in reciprocal space).

6.1. MIR likelihood

In MIR, the data are the observed native and derivative structure-factor amplitudes. Unfortunately, there are significant correlations in the data because all data sets share the scattering from the native protein component, *i.e.* if a reflection is strong/weak in the native then it is likely to be strong/weak in all the derivative data sets as well. To simplify the analysis a (useful) nuisance variable is introduced, the 'true' structure factor $\mathbf{F}_T$, which is the component of scattering shared by the native and derivatives (Read, 2003*a*) and can be thought of as the scattering from a 'true' crystal. With the introduction of $\mathbf{F}_T$ in maximum-likelihood MIR there is nothing 'special' about the native data set. The native is treated in exactly the same way as the derivatives: the native is simply a derivative without heavy atoms. In the nomenclature used here, the observed native and derivative structure factors are all denoted $\mathbf{F}_o$. Elsewhere, $\mathbf{F}_o$ is often written as $\mathbf{F}_P$, denoting that it contains native protein only, or $\mathbf{F}_{PH}$, denoting that it contains native protein and heavy atoms. The data, the set of all 'native' and derivative observed structure-factor amplitudes, is denoted $\{F_{oj}\}$, and the model, the set of all

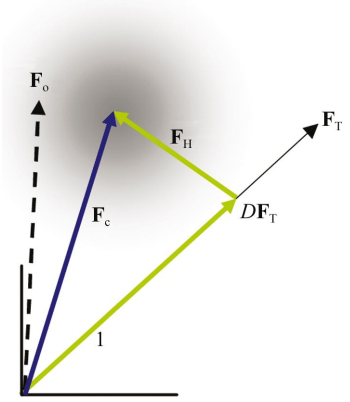

Figure 9
Derivative structure factors. The calculated derivative structure factor ($\mathbf{F}_c$) in blue is the sum of $D\mathbf{F}_T$ and $\mathbf{H}_c$ (in green). The probability distribution for $\mathbf{F}_o$ is shown with grey shading.

calculated heavy-atom structure factors, is denoted $\{\mathbf{F}_{Hj}\}$, with the derivative number denoted by the subscript j. The introduced (useful) nuisance variable $\mathbf{F}_T$ must be integrated out of the probability distribution at the end of the analysis (see §2.6). Since $\mathbf{F}_T$ is a vector, integrating out the parameter requires integrating over the whole complex plane (a double integral, with real and imaginary components integrated from $+\infty$ to $-\infty$). The MIR likelihood function for a reflection is therefore given by

$$P\text{-MIR}_r = P(\text{data};\text{model}) = P(\{F_{oj}\};\{\mathbf{F}_{Hj}\}) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} P(\{F_{oj}\},\mathbf{F}_T;\{\mathbf{F}_{Hj}\})\,\mathrm{d}\mathbf{F}_T. \quad (3)$$

The reason for introducing the nuisance variable $\mathbf{F}_T$ is that by explicitly including the correlated component of the scattering between all the data, the 'leftover' parts of the scattering can be considered to be independent. Therefore, the probabilities for each derivative F_{oj} given its $\mathbf{F}_{Hj}$ and $\mathbf{F}_T$ are (approximately) independent and can be multiplied to give the joint conditional probability (see §2.3),

$$P(\{F_{oj}\};\{\mathbf{F}_{Hj}\},\mathbf{F}_T) = \prod_{j=1}^{N} P(F_{oj};\mathbf{F}_{Hj},\mathbf{F}_T). \quad (4)$$

However, this is an expression for the probability of $\{F_{oj}\}$ given $\{\mathbf{F}_{Hj}\}$ and $\mathbf{F}_T$, not for the probability of $\{F_{oj}\}$ and $\mathbf{F}_T$ given $\{\mathbf{F}_{Hj}\}$, which is what is required for integrating out $\mathbf{F}_T$ (3). The relationship between the two probabilities is given by $P(B,A;C) = P(A;C)\times P(B;C,A)$. Taking $\mathbf{F}_T \equiv \mathrm{A}$, $\{F_{oj}\} \equiv \mathrm{B}$ and $\{\mathbf{F}_{Hj}\} \equiv \mathrm{C}$,

$$P(\{F_{oj}\},\mathbf{F}_T;\{\mathbf{F}_{Hj}\}) = P(\mathbf{F}_T;\{\mathbf{F}_{Hj}\})\times P(\{F_{oj}\};\{\mathbf{F}_{Hj}\},\mathbf{F}_T).$$

If the 'true crystal' lacks atoms at the heavy-atom positions of the derivative, then $P(\mathbf{F}_T;\{\mathbf{F}_{Hj}\})$ is the same as $P(\mathbf{F}_T)$, *i.e.* $\{\mathbf{F}_{Hj}\}$ is irrelevant. $P(\mathbf{F}_T)$ is then the probability of $\mathbf{F}_T$ when all that is known is the number and type of atoms in the 'true crystal' *e.g.* the number of C, N, O and S atoms if the 'true crystal' contains protein only (Wilson, 1949). The probability distribution given by this information is relatively flat and can be ignored (Read, 1991). [However, if $\mathbf{F}_T$ does have atoms coincident with the heavy-atom positions, it should be included (Read, 2003*a*).]

$$P(\{F_{oj}\},\mathbf{F}_T;\{\mathbf{F}_{Hj}\}) = P(\{F_{oj}\};\{\mathbf{F}_{Hj}\},\mathbf{F}_T). \quad (5)$$

Combining (4) and (5),

$$P(\{F_{oj}\},\mathbf{F}_T;\{\mathbf{F}_{Hj}\}) = \prod_{j=1}^{N} P(F_{oj};\mathbf{F}_{Hj},\mathbf{F}_T). \quad (6)$$

Substituting (6) into the integral (3),

$$P\text{-MIR}_r = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\prod_{j=1}^{N} P(F_{oj};\mathbf{F}_{Hj},\mathbf{F}_T)\,\mathrm{d}\mathbf{F}_T. \quad (7)$$

$P(F_{oj};\mathbf{F}_{Hj},\mathbf{F}_T)$ is the probability of the observed structure-factor amplitude F_{oj} given $\mathbf{F}_H$ and $\mathbf{F}_T$ for derivative j; to calculate this probability, $\mathbf{F}_{Hj}$ and $\mathbf{F}_T$ must be used to calculate the structure-factor amplitude F_{cj} that can be compared with F_{oj}. The calculated structure factor (phased) $\mathbf{F}_{cj}$ is the sum of the heavy-atom and protein structure factors (phased) for the derivative. If the heavy-atom model is perfect (and thus $\mathbf{F}_{Hj}$ is perfect) and the protein component of the derivative is identical (isomorphous) to $\mathbf{F}_T$, then the calculated structure factor $\mathbf{F}_{cj}$ is simply given by the sum of $\mathbf{F}_{Hj}$ and $\mathbf{F}_T$. However, $\mathbf{F}_{Hj}$ will not be perfect because the heavy atoms will not have perfect positions and occupancies and some of the sites may be missing from the model and $\mathbf{F}_T$ will not be perfectly isomorphous with the native component of the derivative. Using the same reasoning that applied for developing the refinement target, $\mathbf{F}_T$ and $\mathbf{F}_{Hj}$ are thus down-weighted by D factors ($0 \le D \le 1$). Refining the D factor for $\mathbf{F}_{Hj}$ has the same effect as refining the occupancies and B factors of the heavy atoms and so can be absorbed by these parameters during refinement. Including errors, the calculated structure factor $\mathbf{F}_{cj}$ is given by

$$\mathbf{F}_{cj} = D_j\mathbf{F}_T + \mathbf{F}_{Hj}.$$

The calculated structure-factor amplitude F_{cj} (in terms of $\mathbf{F}_T$ and $\mathbf{F}_{Hj}$) can now be compared with the observed structure-factor amplitude F_{oj},

$$P(F_{oj};\mathbf{F}_{Hj},\mathbf{F}_T) = P(F_{oj};F_{cj}),$$

where $F_{cj} = |D_j\mathbf{F}_T + \mathbf{F}_{Hj}|$.

As was the case for deriving the equation for refinement likelihood and the rotation-function likelihood, the trick to deriving a maximum-likelihood MIR function is to introduce the phase difference α between the observed and calculated structure factors while developing the likelihood function and then to integrate out this (useful) nuisance phase at the end of the analysis (Bricogne, 1991; Read, 1991),

$$P(F_{oj};F_{cj}) = \int_0^{2\pi} P(F_{oj},\alpha;F_{cj})\,\mathrm{d}\alpha.$$

The probability of $\mathbf{F}_{oj}$ is a two-dimensional Gaussian in reciprocal space centred on $\mathbf{F}_{cj}$ with variance $\sigma_{\Delta j}^2$ (Fig. 9),

$$P(\mathbf{F}_{oj};\mathbf{F}_{cj}) = \frac{1}{\pi\sigma_{\Delta j}^2}\exp\left(-\frac{|\mathbf{F}_{oj}-\mathbf{F}_{cj}|^2}{\sigma_{\Delta j}^2}\right).$$

Using this probability and the integral above, it can be shown (Appendix *A*) that the likelihood function is yet another Rice distribution,

$$P(F_{oj};F_{cj}) = \frac{2F_{oj}}{\sigma_{\Delta j}^2}\exp\left(-\frac{F_{oj}^2+F_{cj}^2}{\sigma_{\Delta j}^2}\right)I_0\left(\frac{2F_{oj}F_{cj}}{\sigma_{\Delta j}^2}\right).$$

Experimental errors are incorporated by inflating the variance of the distribution

$$P(F_{oj};F_{cj}) = \frac{2F_{oj}}{\sigma_{\Delta j}^2+\sigma_{Fj}^2}\exp\left(-\frac{F_{oj}^2+F_{cj}^2}{\sigma_{\Delta j}^2+\sigma_{Fj}^2}\right)I_0\left(\frac{2F_{oj}F_{cj}}{\sigma_{\Delta j}^2+\sigma_{Fj}^2}\right). \quad (8)$$

This is the likelihood function for a single reflection and a single derivative. Notice the similarities between this equation and the equations for $P\text{-Xray}_r$ [and $P\text{-TF}_r$; (1)] and $P\text{-RF}_r$ (2). The likelihood function is virtually identical to that for $P\text{-Xray}_r$ except that F_c is not calculated directly from the model but from the sum of $D\mathbf{F}_T$ and $\mathbf{F}_H$. To combine all the

derivatives, the product over all the derivatives is taken before integrating out the nuisance parameter $\mathbf{F}_\mathrm{T}$. Substituting (8) into (7),

$$P\text{-MIR}_r = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\prod_{j=1}^{N}\frac{2F_{\mathrm{o}j}}{\sigma_{\Delta j}^2+\sigma_{Fj}^2}\exp\left(-\frac{F_{\mathrm{o}j}^2+F_{\mathrm{c}j}^2}{\sigma_{\Delta j}^2+\sigma_{Fj}^2}\right)I_0\left(\frac{2F_{\mathrm{o}j}F_{\mathrm{c}j}}{\sigma_{\Delta j}^2+\sigma_{Fj}^2}\right)\mathrm{d}\mathbf{F}_\mathrm{T},$$

where $F_{\mathrm{c}j} = |D_j\mathbf{F}_\mathrm{T} + \mathbf{F}_{\mathrm{H}j}|$.

Unfortunately, the integrating out of $\mathbf{F}_\mathrm{T}$ cannot be performed analytically; it must be performed numerically (values calculated and summed). Double numerical integrations are generally slow to compute and so they have to be performed cleverly.

The MIR likelihood function assumes that errors in the models of heavy atoms are uncorrelated to one another. It also assumes that the non-isomorphism differences between the derivatives are uncorrelated to one another. This is not always the case, particularly when the heavy-atom compounds are chemically related.

6.2. MIRAS likelihood

The likelihood function for MIRAS is the probability of all the F_o^+ and F_o^- given all the calculated heavy-atom structure factors $\mathbf{F}_\mathrm{H}^+$ and $\mathbf{F}_\mathrm{H}^-$ (rather than just the mean F_o and mean $\mathbf{F}_\mathrm{H}$ as for MIR). However, this probability function is difficult to generate by maximum likelihood because F_o^+ and F_o^- are highly correlated (if F_o^+ is large/small then F_o^- will also be large/small). This problem is partially avoided if the mean F_o and anomalous difference ΔF are used instead of F_o^+ and F_o^-, as these are less correlated with one another (if the mean F is large, the anomalous difference ΔF need not be large; North, 1965; Matthews, 1966; de La Fortelle & Bricogne, 1997). The probabilities for the normal and anomalous scattering components are then considered to be independent. The probability of the normal scattering component is the same as that derived for MIR. The probability for the anomalous difference is approximated by a least-squares term (rather than being given by a true likelihood term).

6.3. MAD likelihood

Currently, MAD phasing is treated as a case of MIRAS (de La Fortelle & Bricogne, 1997), where the derivatives correspond to the wavelengths. This is unsatisfactory because the assumption that the errors in the models of heavy atoms between derivatives (*i.e.* wavelengths) are uncorrelated with one another is necessarily violated in MAD. To be treated properly, the likelihood function would have to be computed by performing an integration for each unknown phase. For example, in two-wavelength MAD, four integrations would be required, one each for $\alpha_{\lambda 1}^+$, $\alpha_{\lambda 1}^-$, $\alpha_{\lambda 2}^+$ and $\alpha_{\lambda 2}^-$. Only one such integral can be performed analytically (to give a Rice distribution) and all the others must be performed numerically. Numerical instability and limitations in computing power currently preclude this approach, although Bricogne (2000) has proposed an alternative solution to the problem of performing multiple integrations.

6.4. SAD likelihood

In the special case of SAD, there is a likelihood function that explicitly accounts for the correlations between F_o^+ and F_o^- (Pannu & Read, 2004). The function includes the familiar Rice distribution, which primarily accounts for the anomalous difference, but also another term that accounts for the heavy atoms being part of the model of the normal scatterers (McCoy *et al.*, 2004). Only a single numerical (phase) integration is required. The SAD likelihood for a reflection (P-SAD$_r$) is given by

$$P\text{-SAD}_r = \frac{F_o^-}{\pi\Sigma^-}\int_0^{2\pi}\exp\left(-\frac{|\mathbf{F}_o^- - \mathbf{F}_H^-|^2}{\Sigma^-}\right)\Re(F_o^+, F_c^+, \Sigma^+)\,\mathrm{d}\alpha^-,$$

where $F_c^+ = |\mathbf{F}_H^+ + \mathbf{D}_\Phi(\mathbf{F}_o^- - \mathbf{F}_H^-)|$ and

$$\Re(F_o^+, F_c^+, \Sigma^+) = \frac{2F_o^+}{\Sigma^+}\exp\left(-\frac{F_o^{+2}+F_c^{+2}}{\Sigma^+}\right)I_0\left(\frac{2F_o^+F_c^+}{\Sigma^+}\right).$$

7. Discussion

The Rice distribution is ubiquitous where maximum likelihood is applied in crystallography because it is the result of integrating out the phase of a two-dimensional Gaussian: the phase must be integrated out because only structure-factor amplitudes are measured and two-dimensional Gaussians are ubiquitous because of the central limit theorem or 'random walks' of structure-factor components. In fact, the two-dimensional Gaussians arising from 'random walks' are also fundamentally a result of the central limit theorem. Understanding how and why the Rice distribution arises are concepts that link maximum likelihood to all aspects of macromolecular crystallography.

I hope that this material will give students the confidence to look deeper into the maximum-likelihood literature and discover some of the subtleties lost in the simple explanations. For those who are inspired to know more, the crystallography course notes at http://www-structmed.cimr.cam.ac.uk (by Randy J. Read, Airlie J. McCoy, Andrew G. W. Leslie and Philip R. Evans) are recommended as an appropriate second step.

APPENDIX *A*
Probability distribution for acentric reflections

The probability distribution for $\mathbf{F}_\mathrm{o}$ given $\mathbf{F}_\mathrm{c}$ is a two-dimensional Gaussian with variance σ_Δ^2 centred on $\mathbf{F}_\mathrm{c}$ (Fig. 10),

$$P(\mathbf{F}_\mathrm{o}; \mathbf{F}_\mathrm{c}) = \frac{1}{\pi\sigma_\Delta^2}\exp\left(-\frac{|\mathbf{F}_\mathrm{o} - \mathbf{F}_\mathrm{c}|^2}{\sigma_\Delta^2}\right).$$

The cosine rule with the phase α between $\mathbf{F}_\mathrm{o}$ and $\mathbf{F}_\mathrm{c}$ gives

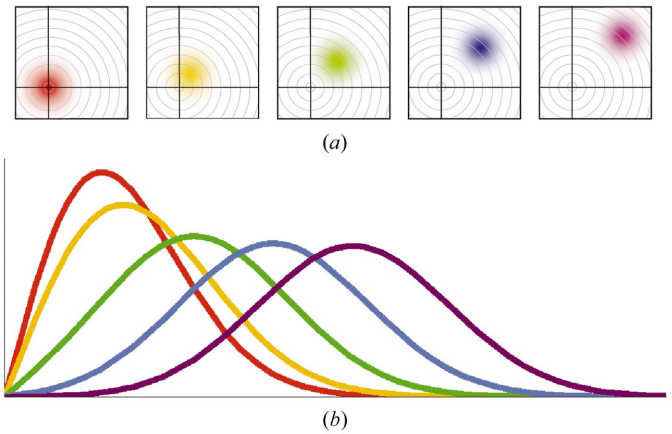

Figure 10
Rice distribution. (*a*) The integral of the two-dimensional Gaussian is the sum of values on concentric circles around the origin (shown in grey). As F_c increases, the two-dimensional Gaussian becomes further offset from the origin (shown with red, yellow, green, blue and purple shading). (*b*) A series of five Rice functions are plotted in red, yellow, green, blue and purple for the different values of F_c indicated with the same colour shading in (*a*). As F_c increases, the Rice function becomes more like a one-dimensional Gaussian.

$$|\mathbf{F}_o - \mathbf{F}_c| = (F_o^2 + F_c^2 - 2F_oF_c\cos\alpha)^{1/2}.$$

The likelihood function is given by integrating out the phase α from the probability $P(F_o, \alpha; F_c)$ (Fig. 10*a*),

$$P(F_o; F_c) = \int_0^{2\pi} P(F_o, \alpha; F_c)\, d\alpha.$$

The relationship between $P(F_o, \alpha; F_c)$ and $P(\mathbf{F}_o; \mathbf{F}_c)$ is given by

$$P(F_o, \alpha; F_c) = F_o \times P(\mathbf{F}_o; \mathbf{F}_c),$$

where the factor F_o is introduced by changing the descriptions of the **F**s from Cartesian coordinates (*i.e.* expressed in terms of real and imaginary components) to polar coordinates (*i.e.* expressed in terms of radial and angular components; this factor is called the Jacobian). Therefore,

$$P(F_o; F_c) = \frac{F_o}{\pi\sigma_\Delta^2}\exp\left(-\frac{F_o^2 + F_c^2}{\sigma_\Delta^2}\right)\int_0^{2\pi}\exp\left(\frac{2F_oF_c}{\sigma_\Delta^2}\cos\alpha\right)\, d\alpha.$$

This integral has an analytical solution of the form

$$\int_0^{2\pi}\exp(z\cos\alpha)\, d\alpha = 2\pi I_0(z),$$

where I_0 is the modified Bessel function of order 0. Therefore,

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

This is known as the Rice distribution (Fig. 10*b*). In the special case where F_c is zero,

$$P(F_o; F_c = 0) = \frac{2F_o}{\sigma_\Delta^2}\exp\left(\frac{F_o^2}{\sigma_\Delta^2}\right).$$

This is known as the Wilson distribution.

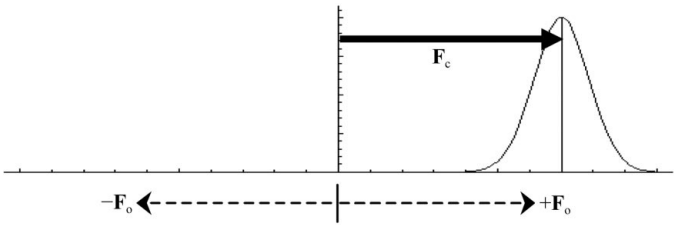

Figure 11
Woolfson distribution. The probability distribution for $\mathbf{F}_o$ given $\mathbf{F}_c$ is a one-dimensional Gaussian centred on $\mathbf{F}_c$. Since the reflection is centric, $\mathbf{F}_o$ either has the same phase as $\mathbf{F}_c$ or is 180° out of phase with $\mathbf{F}_c$.

APPENDIX *B*
Probability distribution for centric reflections

The probability distribution for $\mathbf{F}_o$ given $\mathbf{F}_c$ is a one-dimensional Gaussian with variance σ_Δ^2 centred on $\mathbf{F}_c$. $\mathbf{F}_o$ is either in phase or out of phase with $\mathbf{F}_c$ (Fig. 11). Summing over the two possibilities for the unknown phase,

$$P(F_o; F_c) = \frac{1}{(2\pi\sigma_\Delta^2)^{1/2}}\exp\left[-\frac{(F_o - F_c)^2}{2\sigma_\Delta^2}\right] + \frac{1}{(2\pi\sigma_\Delta^2)^{1/2}}\exp\left[-\frac{(F_o + F_c)^2}{2\sigma_\Delta^2}\right].$$

Expanding the quadratics and using

$$2\cosh x = \exp x + \exp(-x)$$

gives

$$P(F_o; F_c) = \left(\frac{2}{\pi\sigma_\Delta^2}\right)^{1/2}\exp\left(-\frac{F_o^2 + F_c^2}{2\sigma_\Delta^2}\right)\cosh\left(\frac{F_oF_c}{\sigma_\Delta^2}\right).$$

This is known as the Woolfson distribution (Woolfson, 1956). In the special case where F_c is zero,

$$P(F_o; F_c = 0) = \left(\frac{2}{\pi\sigma_\Delta^2}\right)^{1/2}\exp\left(-\frac{F_o^2}{2\sigma_\Delta^2}\right).$$

I thank Randy Read, Garib Murshudov and Laurent Storoni for many useful discussions. I also thank Eleanor Dodson for comments on the manuscript.

References

Bricogne, G. (1991). *Proceedings of the CCP4 Study Weekend. Isomorphous Replacement and Anomalous Scattering*, edited by P. R. Evans & A. G. W. Leslie, pp. 60–68. Warrington: Daresbury Laboratory.
Bricogne, G. (1992). *Proceedings of the CCP4 Study Weekend. Molecular Replacement*, edited by W. Wolf, E. J. Dodson & S. Glover, pp. 62–75. Warrington: Daresbury Laboratory.
Bricogne, G. (1993). *Acta Cryst.* D**49**, 37–60.
Bricogne, G. (2000). *Advanced Special Functions and Applications: Proceedings of the Melfi School on Advanced Topics in Mathematics and Physics*, edited by D. Cocoliccio, G. Dattoli & H. M. Srivastava, pp. 315–232. Rome: Aracne Editrice.

Bricogne, G. & Irwin, J. (1996). *Proceedings of the CCP4 Study Weekend. Macromolecular Refinement*, edited by E. Dodson. M. Moore, A. Ralph & S. Bailey, pp. 85–92. Warrington: Daresbury Laboratory.
Engh, R. A. & Huber, R. (1991). *Acta Cryst.* A**47**, 392–400.
Frieden, B. R. (1985). *J. Opt. Soc. Am.* **73**, 1764–1770.
Green, E. A. (1979). *Acta Cryst.* A**35**, 351–359.
Hendrickson, W. A. & Lattman, E. E. (1970). *Acta Cryst.* B**26**, 136–143.
Jaynes, E. T. (1968). *IEEE Trans. Syst. Sci. Cybern.* **SSC-4**, 227–241.
Jaynes, E. T. (1979). *The Maximum Entropy Formalism*, edited by R. D. Levine & M. Tribus, pp 15–118. Cambridge: MIT Press.
La Fortelle, E. de & Bricogne, G. (1997). *Methods Enzymol.* **276**, 472–494.
McCoy, A. J. (2002). *Curr. Opin. Struct. Biol.* **12**, 670–673.
McCoy, A. J., Storoni, L. C. & Read, R. J. (2004). *Acta Cryst.* D**60**, 1220–1228.
Matthews, B. W. (1966). *Acta Cryst.* **20**, 82–86.
Mohammed-Djafari, A. (2003). *Am. Inst. Phys. Conf. Proc.* **659**, 281–306.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst* D**53**, 240–255.
North, A. C. T. (1965). *Acta Cryst.* **18**, 212–216.
Pannu, N. S., Murshudov, G. N., Dodson, E. J. & Read, R. J. (1998). *Acta Cryst.* D**54**, 1285–1294.
Pannu, N. S. & Read, R. J. (1996). *Acta Cryst.* A**52**, 659–668.
Pannu, N. S. & Read, R. J. (2004). *Acta Cryst.* D**60**, 22–27.
Read, R. J. (1990). *Acta Cryst.* A**46**, 900–912.
Read, R. J. (1991). *Proceedings of the CCP4 Study Weekend. Isomorphous Replacement and Anomalous Scattering*, edited by P. R. Evans & A. G. W. Leslie, pp. 69–79. Warrington: Daresbury Laboratory.
Read, R. J. (1994). *Lecture Notes from the Workshop on Isomorphous Replacement Methods in Macromolecular Crystallography*. Am. Crystallogr. Assoc. Ann. Meet., Atlanta, GA, USA.
Read, R. J. (2001). *Acta Cryst.* D**57**, 1373–1382.
Read, R. J. (2003*a*). *Acta Cryst.* D**59**, 1891–1902.
Read, R. J. (2003*b*). *Crystallogr. Rev.* **9**, 33–41.
Sim, G. A. (1959). *Acta Cryst.* **12**, 813–815.
Srinivasan, R. & Ramachandran, G. N. (1965). *Acta Cryst.* **19**, 1003–1007.
Storoni, L. C., McCoy, A. J. & Read, R. J. (2004). *Acta Cryst.* D**60**, 432–438.
Terwilliger, T. C. (2000). *Acta Cryst.* D**56**, 965–972.
Wilson, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.
Woolfson, M. M. (1956). *Acta Cryst.* **9**, 804–810.