Linking Crystallographic Model and Data Quality

P. Andrew Karplus and Kay Diederichs

Science **336**, 1030 (2012);
DOI: 10.1126/science.1218231

(*15*) [rockrose averages 23 times greater ground cover than dove's-foot cranesbill per 100 × 100 m sample area that contains host plants (fig. S1) (*11*)]. The long-lived perennial rockrose also has more stable populations than does the annual and ruderal dove's-foot cranesbill. These differences between the plants enable rockrose to support larger (Fig. 1, C and D) and more stable brown argus populations [Levene's test for equality of variances: $W = 10.97$, $n = 2$ 30-generation sequences, $P = 0.002$ (Fig. 1D)]. Moreover, rockrose frequently grows in areas of short turf on southerly facing slopes (fig. S2) (*11*, *20*), which provide warm microclimates [southerly aspects receive greater direct radiation and achieve higher maximum summer temperatures (*21*)]. As recently as the early 1980s, the brown argus was mainly associated with rockrose populations on sheltered south-facing slopes (*12*). The few historical records of Geraniaceae-feeding populations from this period were predominantly in sand dunes (*13*), which also provide warm microclimates.

Summer temperatures in Britain from 1990 to 2009 were on average 0.78°C warmer than between 1800 and 1989, and this is likely to have increased the thermal suitability of sites for brown argus, especially those that are not southerly facing. This would have increased the ability of Geraniaceae-containing sites to support brown argus population growth; there was a 5.3-fold increase in brown argus population density in Geraniaceae sites between 1976–1985 and 2000–2009 (Spearman's rank correlation between year and density on Geraniaceae: $r_s = 0.76$, $n = 34$ years, $P < 0.001$) (Fig. 1C). In contrast, no increase in overall population density occurred at rockrose sites (Spearman's rank correlation between year and density on rockrose: $r_s = 0.25$, $n = 34$ years, $P = 0.162$; 1.1-fold density increase from 1976–1985 to 2000–2009) (Fig. 1C), even though butterfly abundance increased temporarily during warm summers. This suggests that other factors limit population density on rockrose (supplementary text).

Based on 100 × 100 m grid squares with records of host plants, dove's-foot cranesbill is 4 to 17 times more widespread than is rockrose in counties where rapid expansion has taken place (Fig. 3) (*11*). Once the brown argus can establish populations on cranesbill, the high frequency of cranesbill populations in the landscape permit it to spread between populations of this host plant without the need for long-distance dispersal. The butterfly's capacity to use Geranicaeae has been aided by the spread of butterfly phenotypes that readily select Geraniaceae plants for egg-laying (*15*) and by a degree of escape from natural enemies (parasitoids) associated with historical rockrose sites (*22*). These processes have come together to generate an unexpectedly rapid transformation in the metapopulation dynamics of the butterfly from a highly localized distribution associated with southerly facing rockrose-containing calcareous grasslands to widespread use of virtually any grassland with rockrose or Geranicaeae host plants. Ecological and evolutionary adjustments by the butterfly, interacting with alternative host plants that differ in their niches and life-history traits, have resulted in rapid range expansion of this previously rare and declining butterfly. We suggest that altered interactions among species do not necessarily constrain distribution changes but can facilitate expansions.

References and Notes

1. R. Hickling *et al.*, *Glob. Change Biol.* **12**, 450 (2006).
2. M. S. Warren *et al.*, *Nature* **414**, 65 (2001).
3. I. C. Chen *et al.*, *Science* **333**, 1024 (2011).
4. J. K. Hill *et al.*, *Ecol. Lett.* **4**, 313 (2001).
5. T. Park, *Physiol. Zool.* **27**, 177 (1954).
6. A. J. Davis *et al.*, *Nature* **391**, 783 (1998).
7. A. J. Davis *et al.*, *J. Anim. Ecol.* **67**, 600 (1998).
8. W. H. Van der Putten, M. Macel, M. E. Visser, *Philos. Trans. R. Soc. London Ser. B* **365**, 2025 (2010).
9. M. B. Araújo, M. Luoto, *Glob. Ecol. Biogeogr.* **16**, 743 (2007).
10. O. Schweiger *et al.*, *Ecology* **89**, 3472 (2008).
11. Materials and methods are available as supplementary materials on *Science* Online.
12. N. A. D. Bourn, J. A. Thomas, *Biol. Conserv.* **63**, 67 (1993).
13. J. Heath, E. Pollard, J. A. Thomas, *Atlas of Butterflies in Britain and Ireland* (Viking, Harmondsworth, 1984).
14. J. Asher *et al.*, *The Millennium Atlas of Butterflies in Britain and Ireland* (Oxford Univ. Press, Oxford, 2001).
15. C. D. Thomas *et al.*, *Nature* **411**, 577 (2001).
16. T. Tolman, *Collins Field Guide to the Butterflies of Britain and Europe* (HarperCollins, London, 1997).
17. E. Pollard, T. J. Yates, *Monitoring Butterflies for Ecology and Conservation* (Chapman & Hall, London, 1993).
18. M. C. F. Proctor, M. E. Griffiths, *J. Ecol.* **44**, 675 (1956).
19. J. P. Grime, J. G. Hodgson, R. Hunt, *Comparative Plant Ecology: A Functional Approach to Common British Species* (Unwin Hyman, London, 1988).
20. K. H. Lakhani, B. N. K. Davis, *J. Appl. Ecol.* **19**, 621 (1982).
21. R. B. Hutchins *et al.*, *Soil Sci.* **121**, 234 (1976).
22. R. Menéndez *et al.*, *Ecol. Entomol.* **33**, 413 (2008).

Acknowledgments: Work was funded through NERC Ecology and Hydrology Funding Initiative grant NE/E012035. Butterfly distribution data were derived from a database of records submitted by volunteers (www.butterfly-conservation.org/text/64/butterfly_distribution.html), and densities were derived from UK Butterfly Monitoring Scheme data (www.ukbms.org/obtaining.htm). Both schemes are operated by Butterfly Conservation and the Centre for Ecology & Hydrology and funded by a consortium of government agencies. Plant data were derived from a database of volunteer records managed by the Botanical Society for the British Isles (www.bsbi.org.uk/research.html). We are grateful to the volunteers who collected the original butterfly and plant data. Temperature data were derived from the Central England Temperature (CET) data set (www.metoffice.gov.uk/hadobs). Digital Elevation Models were provided by the NERC Earth Observation Data Centre (www.neodc.rl.ac.uk/browse/neodc/nextmap). Original data collected by authors is presented in the supplementary materials.

Supplementary Materials

www.sciencemag.org/cgi/content/full/336/6084/1028/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S4
Tables S1 and S2
References (*23*–*28*)

22 November 2011; accepted 6 April 2012
10.1126/science.1216980

Linking Crystallographic Model and Data Quality

P. Andrew Karplus[1] and Kay Diederichs[2]*

In macromolecular x-ray crystallography, refinement R values measure the agreement between observed and calculated data. Analogously, R_{merge} values reporting on the agreement between multiple measurements of a given reflection are used to assess data quality. Here, we show that despite their widespread use, R_{merge} values are poorly suited for determining the high-resolution limit and that current standard protocols discard much useful data. We introduce a statistic that estimates the correlation of an observed data set with the underlying (not measurable) true signal; this quantity, CC*, provides a single statistically valid guide for deciding which data are useful. CC* also can be used to assess model and data quality on the same scale, and this reveals when data quality is limiting model improvement.

Accurately determined protein structures provide insight into how biology functions at the molecular level and also guide the development of new drugs and protein-based nanomachines and technologies. The large majority of protein structures are determined by x-ray crystallography, where measured diffraction data are used to derive a molecular model. It is surprising that, despite decades of methodology development, the question of how to select the resolution cutoff of a crystallographic data set is still controversial, and the link between the quality of the data and the quality of the derived molecular model is poorly understood. Here, we describe a statistical quantity that addresses both of these issues and will lead to improved molecular models.

The measured data in x-ray crystallography are the intensities of reflections, and these yield structure factor amplitudes each with unique h, k, and l indices that define the lattice planes. The standard indicator for assessing the agreement of

[1]Department of Biochemistry and Biophysics, Oregon State University, Corvallis, OR 97331, USA. [2]Department of Biology, University of Konstanz, Box 647, D-78457 Konstanz, Germany.

*To whom correspondence should be addressed. E-mail: kay.diederichs@uni-konstanz.de

a refined model with the data is the crystallographic R value, defined as

$$R = \frac{\sum_{hkl} |F_{obs}(hkl) - F_{calc}(hkl)|}{\sum_{hkl} F_{obs}(hkl)} \quad (1)$$

where $F_{obs}(hkl)$ and $F_{calc}(hkl)$ are the observed and calculated structure factor amplitudes, respectively. R is 0.0 for perfect agreement with the data, and R is near 0.59 for a random model (*1*). Because R can be made arbitrarily low for models having sufficient parameters to overfit the data, Brünger (*2*) introduced R_{free} as a cross-validated R on the basis of a small subset of reflections not used during refinement. The R for the larger "working" set of reflections is then referred to as R_{work}.

Crystallographic data quality is commonly assessed by an analogous indicator R_{merge} [originally (*3*) R_{sym}], which measures the spread of n independent measurements of the intensity of a reflection, $I_i(hkl)$, around their average, $\bar{I}(hkl)$:

$$R_{merge} = \frac{\sum_{hkl} \sum_{i=1}^{n} |I_i(hkl) - \bar{I}(hkl)|}{\sum_{hkl} \sum_{i=1}^{n} I_i(hkl)} \quad (2)$$

In 1997, it was discovered that because $I_i(hkl)$ values influence $\bar{I}(hkl)$, the R_{merge} definition must be adjusted by a factor of $\sqrt{n/(n-1)}$ to give values that are independent of the multiplicity (*4*). The multiplicity-corrected version, called R_{meas}, reliably reports on the consistency of the individual measurements. A further variant, R_{pim} (*5*), reports on the expected precision of $\bar{I}(hkl)$ and is lower by a factor of $1/\sqrt{n}$ factor compared with R_{meas}. Because the strength of diffraction decreases with resolution, a high-resolution cutoff is applied to discard data considered so noisy that their inclusion might degrade the quality of the resulting model. Data are typically truncated at a resolution before the R_{merge} (or R_{meas}) value exceeds ~0.6 to 0.8 and before the empirical signal-to-noise ratio, $\langle \bar{I}/\sigma(\bar{I}) \rangle$, drops below ~2.0 (*6*) (fig. S1). The uncertainty associated with these criteria is illustrated by a recent review that concluded "an appropriate choice of resolution cutoff is difficult and sometimes seems to be performed mainly to satisfy referees" (*6*).

That these criteria result in high-resolution cutoffs that are too conservative is illustrated here using an example data set (***EXP***) collected for a cysteine-bound complex of cysteine dioxygenase (CDO); the ***EXP*** data have an average intensity about 7% as strong as the data originally used to determine the structure at 1.42 Å resolution (PDB 3ELN; R_{work}/R_{free} = 0.135/0.177) (*7*, *8*). Standardized model refinements starting with a 1.5 Å resolution unliganded CDO structure (PDB code 2B5H) (*9*) were carried out against the ***EXP*** data for a series of high-resolution cutoffs between 2.0 and 1.42 Å resolution (table S1). As R value comparisons are only meaningful if calculated at the same resolution, we evaluated paired refinements made with adjacent resolution limits using R_{work} and R_{free} values calculated at the poorer resolution limit. Improvement is indicated by drops in R_{free} or increases in R_{work} at the same R_{free} (meaning the model is less overfit). This analysis revealed that every step of added data improved the resulting model (Fig. 1). Consistent with this, difference Fourier maps show a similar trend in signal versus resolution (fig. S2), and geometric parameters of the resulting models improve with resolution (table S2).

The proven value of the data out to 1.42 Å resolution contrasts strongly with the R_{meas} and $\langle \bar{I}/\sigma(\bar{I}) \rangle$ values at that resolution (>4.0 and ~0.3, respectively) (Fig. 2), which are far beyond the limits currently associated with useful data. Applying the typical standards described above, this data set would have been truncated at ~1.8 Å resolution, which would halve the number of unique reflections in the data set (table S1) and would yield a worse model.

It is striking to observe the different behavior at high resolution of the crystallographic versus the data-quality R values, with the one remaining below 0.40 and the other diverging toward infinity (Fig. 2). Consideration of the R_{merge} formula rationalizes this divergence, because the denominator (the average net intensity) approaches zero at high resolution, but the numerator becomes dominated by background noise and is essentially constant. Thus, despite their similar names and mathematical definitions, data-quality

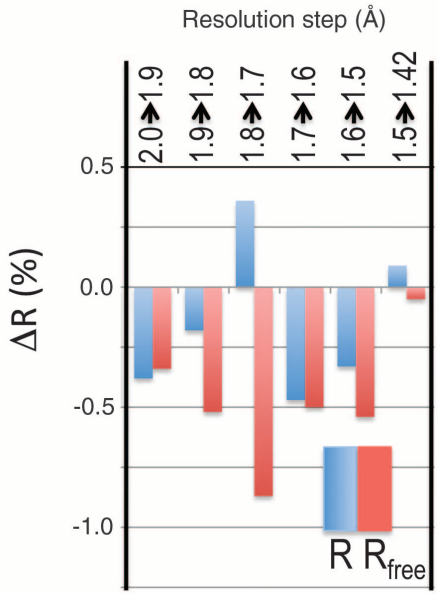

Fig. 1. Higher-resolution data, even if weak, improves refinement behavior. For each incremental step of resolution from $X \rightarrow Y$ (top legend), the pair of bars gives the changes in overall R_{work} (blue) and R_{free} (red) for the model refined at resolution Y with respect to those for the model refined at resolution X, with both R values calculated at resolution X. The first pair of bars shows that R_{work} and R_{free} dropped 0.38% and 0.34%, respectively, upon isotropic refinement when the refinement resolution limit was extended from 2.0 to 1.9 Å; the other pairs of bars show the improvement upon anisotropic refinement.

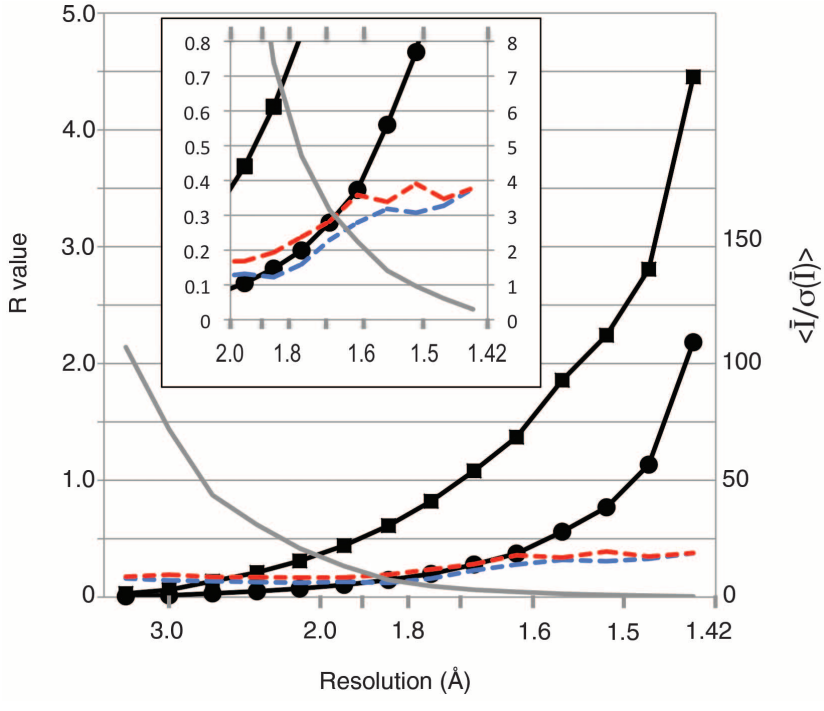

Fig. 2. Data quality R values behave differently than those from crystallographic refinement, and useful data extend well beyond what standard cutoff criteria would suggest. R_{meas} (squares) and R_{pim} (circles) are compared with R_{work} (blue) and R_{free} (red) from 1.42 Å resolution refinements against the ***EXP*** data set. $\langle \bar{I}/\sigma(\bar{I}) \rangle$ (gray) is also plotted. **(Inset)** A close-up of the plot beyond 2 Å resolution.

R values are not comparable to R values from model refinement, and there is no valid basis for the commonly applied criterion that data are not useful beyond a resolution where R_{meas} (or R_{merge} or R_{pim}) rises above ~0.6. As suggested by Wang (*10*), $\langle \bar{I}/\sigma(\bar{I}) \rangle$ at a much lower level than generally recommended could be used to define the cutoff, but this has the problem that $\sigma(\bar{I})$ values can be misestimated (*6*, *11*).

With current standards not serving as reliable guides for selecting a high-resolution cutoff, we investigated the use of the Pearson correlation coefficient (CC) (*12*) as a parameter that could potentially assess both data accuracy and the agreement of model and data on a common scale. Pearson's CC is already used in crystallography, in that a CC value of 0.3 between independent measurements of anomalous signals has become the recommended criterion for selecting the high-resolution cutoff of the data to be used for defining the locations of the anomalous scatterers (*13*). Following a procedure suggested earlier (*4*), we divided the unmerged ***EXP*** data into two parts, each containing a random half of the measurements of each unique reflection. Then, the CC was calculated between the average intensities of each subset. This quantity, denoted $CC_{1/2}$, is near 1.0 at low resolution and drops to near 0.1 at high resolution (Fig. 3). According to Student's t test (*12*), the $CC_{1/2}$ of 0.09 for the ~2100 reflection pairs in the highest resolution bin is significantly different from zero ($P = 2 \times 10^{-5}$).

This high significance occurs even though $CC_{1/2}$ should be expected to underestimate the information content of the data. This is because for weak data, $CC_{1/2}$ measures the correlation of one noisy data set (the first half-data set) with another noisy data set (the other half-data set), whereas the true level of signal would be measured by what could be called CC_{true}, the correlation of the averaged data set (less noisy because of the extra averaging) with the noise-free true signal. Although the true signal would normally not be known, for the ***EXP*** test case, the 3ELN data provide a reference that has much lower noise and should be much closer to the underlying true data. The CC calculated between the ***EXP*** and 3ELN data sets is indeed uniformly higher than $CC_{1/2}$ (Fig. 3), dropping only to 0.31 in the highest resolution bin (Student's t test $P = 10^{-64}$).

We next sought an analytical relation between $CC_{1/2}$ and CC_{true}. Using only the assumption that errors in the two half-data sets are random and, on average, of similar size (see supplementary text), we derived the relation

$$CC^* = \sqrt{\frac{2CC_{1/2}}{1 + CC_{1/2}}} \qquad (3)$$

where CC* estimates the value of CC_{true}, based on a finite-size sample. Equation 3 has been used in electron microscopy studies for a similar purpose (*14*) and is also related to the Spearman-Brown prophecy formula used in psychometrics to predict what test length is required to achieve a certain level of reliability (*15*). CC*, when computed with Eq. 3, agrees reasonably well with the CC for the ***EXP*** data compared with the 3ELN reference data, which shows that systematic factors influencing a real data set are not large enough to greatly perturb this relation (Fig. 4A). CC* provides a statistic that not only assesses data quality but also allows direct comparison of crystallographic model quality and data quality on the same scale. In particular, CC_{work} and CC_{free}—the standard and cross-validated correlations of the experimental intensities, with the intensities calculated from the refined molecular model—can be directly compared with CC* (Fig. 4B). A CC_{work} larger than CC* implies overfitting, because, in that case, the model agrees better with the experimental data than the true signal does. A CC_{free} smaller than CC* (such as is seen at low resolution) indicates that the model does not account for all of the signal in the data. A CC_{free} closely matching CC*, such as at high resolution in Fig. 4B, implies that data quality is limiting model improvement. In this high-resolution region, the model, which was refined against ***EXP***, correlates much better with the more accurate 3ELN than with the ***EXP*** data (Fig. 4B). This shows that, as is common for parsimonious models (*16*), the constructed molecular model is a better predictor of the true signal than are the experimental data from which it was derived. On a related point, because current estimates of a model's coordinate error do not take the data errors into account (*17*–*19*), the model accuracy is actually better than these methods indicate.

We verified, using a simulated data set (*20*) and two further test cases, that these findings are

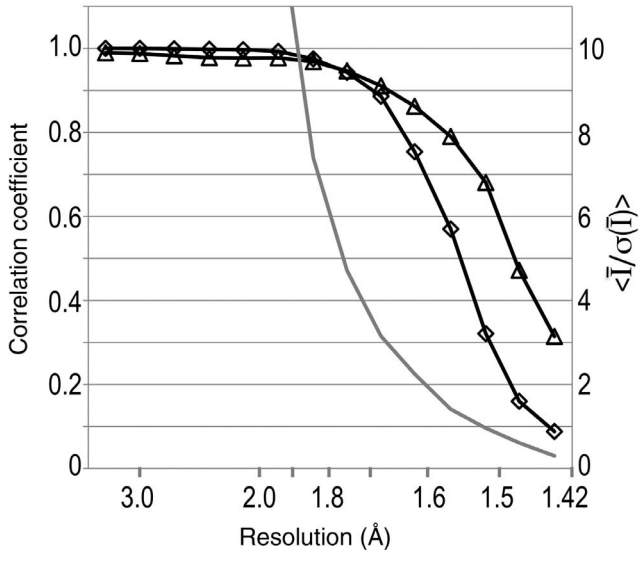

Fig. 3. Signal as a function of resolution as measured by correlation coefficients. Plotted as a function of resolution for the ***EXP*** data are $CC_{1/2}$ (diamonds) and the CC for a comparison with the 3ELN reference data set (triangles). $\langle \bar{I}/\sigma(\bar{I}) \rangle$ (gray) is also shown. All determined $CC_{1/2}$ values shown have expected standard errors of <0.025 (*21*, *22*).

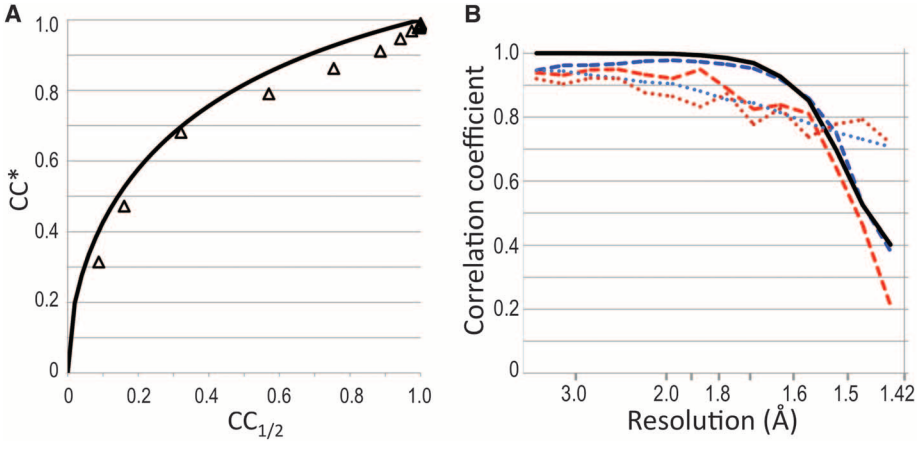

Fig. 4. The $CC_{1/2}$/CC* relation and the utility of comparing CC* with CC_{work} and CC_{free} from a refined model. (**A**) Plotted is the analytical relation (Eq. 3) between $CC_{1/2}$ and CC* (black curve). Also roughly following the CC* curve are the CC values for the ***EXP*** data compared with 3ELN (triangles) as a function of $CC_{1/2}$. (**B**) Plotted as a function of resolution are CC* (black solid) for the ***EXP*** data set, as well as CC_{work} (blue dashed) and CC_{free} (red dashed) calculated on intensities from the 1.42 Å refined model. Also shown are values for CC_{work} (blue dotted) and CC_{free} (red dotted) between the 1.42 Å refined model and the 3ELN data set.

not specific to the ***EXP*** data (tables S3, S4, and S5, and fig. S3). Thus, CC* (or $CC_{1/2}$) is a robust, statistically informative quantity useful for defining the high-resolution cutoff in crystallography. These examples show that with current data reduction and refinement protocols, it is justified to include data out to well beyond currently employed cutoff criteria (fig. S4), because the data at these lower signal levels do not degrade the model, but actually improve it. Advances in data-processing and refinement procedures, which until now have not been optimized for handling such weak data, may lead to further improvements in model accuracy. Finally, we emphasize that the analytical relation (Eq. 3) between $CC_{1/2}$ and CC* is general, and thus, CC* may have similar applications for data- and model-quality assessment in other fields of science involving multiply measured data.

References and Notes

1. A. J. C. Wilson, *Acta Crystallogr.* **3**, 397 (1950).
2. A. T. B. Brünger, *Nature* **355**, 472 (1992).
3. U. W. Arndt, R. A. Crowther, J. F. W. Mallett, *J. Phys. E Sci. Instrum.* **1**, 510 (1968).
4. K. Diederichs, P. A. Karplus, *Nat. Struct. Biol.* **4**, 269 (1997).
5. M. S. Weiss, *J. Appl. Cryst.* **34**, 130 (2001).
6. P. R. Evans, *Acta Crystallogr. D Biol. Crystallogr.* **67**, 282 (2011).
7. C. R. Simmons *et al.*, *Biochemistry* **47**, 11390 (2008).
8. Materials and methods are available as supplementary materials on *Science* Online.
9. C. R. Simmons *et al.*, *J. Biol. Chem.* **281**, 18723 (2006).
10. J. Wang, *Acta Crystallogr. D Biol. Crystallogr.* **66**, 988 (2010).
11. P. Evans, *Acta Crystallogr. D Biol. Crystallogr.* **62**, 72 (2006).
12. N. A. Rahman, *A Course in Theoretical Statistics* (Griffin, London, 1968).
13. T. R. Schneider, G. M. Sheldrick, *Acta Crystallogr. D Biol. Crystallogr.* **58**, 1772 (2002).
14. P. B. Rosenthal, R. Henderson, *J. Mol. Biol.* **333**, 721 (2003).
15. P. Bobko, *Correlation and Regression: Applications for Industrial Organizational Psychology and Management* (Sage Publications, Thousand Oaks, 2001).
16. H. G. Gauch Jr., *Am. Sci.* **81**, 468 (1993).
17. V. Luzzati, *Acta Crystallogr.* **6**, 142 (1953).
18. D. W. J. Cruickshank, *Acta Crystallogr. D Biol. Crystallogr.* **55**, 583 (1999).
19. R. A. Steiner, A. A. Lebedev, G. N. Murshudov, *Acta Crystallogr. D Biol. Crystallogr.* **59**, 2114 (2003).
20. K. Diederichs, *Acta Crystallogr. D Biol. Crystallogr.* **65**, 535 (2009).
21. R. Fisher, *Statistical Methods for Research Workers* (Oliver and Boyd, Edinburgh, 1925), §33–34.
22. Ten independent random partitionings of the data into the two subsets for calculating $CC_{1/2}$ yielded standard deviations of <0.02 in all resolution ranges, and agreed reasonably with the expected standard error as calculated by $\sigma(CC) = (1 - CC^2)/\sqrt{n - 1}$ where n is the number of observations contributing to the CC calculation (*21*).

Acknowledgments: This work was supported in part by the Alexander von Humboldt Foundation, the Konstanz Research School Chemical Biology, and NIH grants GM083136 and DK056649. We thank R. Cooley for providing the EXP data images, V. Lunin for help with deriving Eq. 3, and A. Gittleman for help with mathematical notation. We also thank M. Junk, W. Kabsch, K. Schäfer, D. Tronrud, M. Wells and W. Welte for critically reading the manuscript. The program HIRESCUT is available upon request. P.A.K. and K.D. designed and performed the research and wrote the paper. The authors declare no competing financial interests.

Supplementary Materials

www.sciencemag.org/cgi/content/full/336/6084/1030/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S4
Tables S1 to S5
References (*23*–*29*)

20 December 2011; accepted 16 March 2012
10.1126/science.1218231

Structures from Anomalous Diffraction of Native Biological Macromolecules

Qun Liu,[1] Tassadite Dahmane,[2] Zhen Zhang,[2] Zahra Assur,[2] Julia Brasch,[2] Lawrence Shapiro,[2] Filippo Mancia,[3] Wayne A. Hendrickson[1,2,3,4]*

Crystal structure analyses for biological macromolecules without known structural relatives entail solving the crystallographic phase problem. Typical de novo phase evaluations depend on incorporating heavier atoms than those found natively; most commonly, multi- or single-wavelength anomalous diffraction (MAD or SAD) experiments exploit selenomethionyl proteins. Here, we realize routine structure determination using intrinsic anomalous scattering from native macromolecules. We devised robust procedures for enhancing the signal-to-noise ratio in the slight anomalous scattering from generic native structures by combining data measured from multiple crystals at lower-than-usual x-ray energy. Using this multicrystal SAD method (5 to 13 equivalent crystals), we determined structures at modest resolution (2.8 to 2.3 angstroms) for native proteins varying in size (127 to 1148 unique residues) and number of sulfur sites (3 to 28). With no requirement for heavy-atom incorporation, such experiments provide an attractive alternative to selenomethionyl SAD experiments.

Crystallographic structure determinations for biomolecules require the retrieval of phases, which are lost when measuring x-ray diffraction patterns. For the first protein crystal structures, phase evaluation was by the method of multiple isomorphous replacement (MIR) with derivatives incorporating mercury [atomic number (Z) = 80] or other heavy atoms. Once many structures were known, phases could often be estimated by the method of molecular replacement; however, de novo structure determination remained essential for molecules without adequately close structural relatives. Multiwavelength anomalous diffraction (MAD) analyses (*1*), which exploit element-specific scattering from x-ray resonance with atomic orbitals, came to be used increasingly for de novo structures as tunable synchrotron beamlines developed (*2*). Whereas MAD gives definitive phase information, its single-wavelength counterpart, SAD, is ambiguous in defining only trigonometric sines of phases. This phase ambiguity could be resolved once density-modification procedures, based largely on molecular boundaries and symmetry, were devised (*3*, *4*); and SAD then surged (*5*). MAD and SAD now dominate de novo phasing, as they have the advantage that lighter atoms can be effective sources of phasing signals. Selenomethionine is easily incorporated into proteins (*6*), and selenium (Z = 34) is now by far the most-used phasing element (*2*). With MAD and SAD, metal atoms such as iron (Z = 26) present in some native proteins can also suffice.

Sulfur (Z = 16) is the heaviest element in most native proteins. Its K-shell resonance at 2.47 keV (λ = 5.02 Å) is inaccessible to standard MAD experiments, and its anomalous scattering at conventional wavelengths is slight; nevertheless, sulfur anomalous scattering can suffice for SAD phasing. The structure of crambin was the first to be determined from sulfur SAD phasing (*7*), although the experiment was not then identified as SAD. Later, broader effectiveness of sulfur SAD was demonstrated with tests on lysozyme (*8*) and in solving the structure of obelin (*9*). Similarly, the feasibility of phosphorous SAD was demonstrated for nucleic acids (*10*). The motivation for truly routine native SAD is great, because heavy-atom incorporations are often problematic, even for the most reliable selenomethionine. Subsequent optimization of native-SAD experiments has included developments for low-energy measurements (*11*), assessments of the impact of high data redundancy (*10*, *11*), optimal wavelength selection (*12*), control of complications from radiation damage (*13*, *14*), and the use of home-source CrKα radiation (*15*).

Besides test cases and technical developments, some novel protein structures beyond crambin and obelin have been determined by sulfur SAD analyses. As compared with the swelling numbers of SAD structures in general, however, the

[1]New York Structural Biology Center, National Synchrotron Light Source (NSLS) X4, Building 725, Brookhaven National Laboratory, Upton, NY 11973, USA. [2]Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032, USA. [3]Department of Physiology and Cellular Biophysics, Columbia University, New York, NY 10032, USA. [4]Howard Hughes Medical Institute, Columbia University, New York, NY 10032, USA.

*To whom correspondence should be addressed. E-mail: wayne@convex.hhmi.columbia.edu