electronic reprint

Journal of
**Applied Crystallography**
ISSN 0021-8898
Editor: **Gernot Kostorz**

# Automated diffraction image analysis and spot searching for high-throughput crystal screening

**Zepu Zhang, Nicholas K. Sauter, Henry van den Bedem, Gyorgy Snell and Ashley M. Deacon**

Journal of
**Applied Crystallography**
ISSN 0021-8898

Received 1 September 2005
Accepted 5 December 2005



# Automated diffraction image analysis and spot searching for high-throughput crystal screening

**Zepu Zhang,[a] Nicholas K. Sauter,[b] Henry van den Bedem,[c] Gyorgy Snell[d] and Ashley M. Deacon[c]***

[a]CISES, The University of Chicago, 5734 South Ellis Avenue, Chicago, IL 60637, USA, [b]Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, CA 94720, USA, [c]Stanford Synchrotron Radiation Laboratory, 2575 Sand Hill Road, Menlo Park, CA 94025, USA, and [d]Syrrx, Inc., 10410 Science Center Drive, San Diego, CA 92121, USA. Correspondence e-mail: adeacon@slac.stanford.edu

A new software package, *DISTL* (*Diffraction Image Screening Tool and Library*), for the rapid analysis of X-ray diffraction patterns collected from macromolecular crystals is presented. Within seconds, the program characterizes the strength and quality of the Bragg spots, determines the limiting resolution of the image, and identifies deleterious features such as ice-rings and intense salt reflections. The procedure also generates a reliable set of intense peaks for autoindexing. The ability to classify a large number of crystals quickly will be especially useful at synchrotron and home-laboratory X-ray sources where automated crystal screening and data collection systems have been implemented.

## 1. Introduction

Research in macromolecular crystallography is rapidly becoming more 'large scale'. Over the past few years new technology has been developed to facilitate structural genomics research (Stevens *et al.*, 2001). Automated protein production and crystallization facilities are being used to target the protein products of an entire genome in parallel (Lesley *et al.*, 2002). There is also a continual drive within the wider structural biology community toward higher-throughput experiments, including mutational studies of protein and DNA, combinatorial approaches to drug design, and the study of large macromolecular assemblies.

All of these types of experiments can benefit from more systematic and automated procedures for evaluating crystal diffraction. It is important to be able to characterize multiple crystals so that the conditions used for crystal growth and cryoprotection can be optimized. Furthermore, since crystal quality is often highly variable even within batches grown under similar conditions, crystal screening is necessary in order to select the best samples for full data collection.

At synchrotron sources, many beamlines are now equipped with automated sample-handling equipment, which can transfer a cryo-cooled crystal from a storage vessel onto the goniometer stage (Muchmore *et al.*, 2000; Cohen *et al.*, 2002; Snell *et al.*, 2004). Graphical user interfaces, such as *BLU-ICE* (McPhillips *et al.*, 2002; Miller *et al.*, 2004), allow dozens of samples to be targeted for screening. However, each acquired image must still be evaluated by visual inspection, which prohibits the complete automation of the crystal screening process and presents several other significant drawbacks to the high-throughput experiment. Notably, this approach lacks the objectivity provided by quantitative measurement. Important values, such as the number and quality of Bragg spots, the limiting resolution of the diffraction pattern, and the presence of ice-ring artifacts, will vary depending on the experience of the individual user and the amount of time spent evaluating each image. Sometimes important characteristics of the diffraction pattern will be overlooked.

Eliminating the labor-intensive step of visual evaluation has an additional significance beyond just increasing the number of crystals that can be screened. It will make quantitative information about each diffraction image available in 'real time' during data collection. Slowly varying conditions such as unwanted ice build-up and loss of resolution due to radiation damage can be under constant scrutiny. By identifying Bragg spots and separating them from artifacts, the Bragg spot positions become available for automatic indexing, also free from the need for visual inspection. In principle, automated image analysis facilitates a higher level of beamline control, making it possible to solve a structure automatically starting from a set of uncharacterized crystal samples.

To further these objectives we developed *DISTL* (*Diffraction Image Screening Tool and Library*), a C++ class library for image characterization. A stand-alone C++ program, *Spotfinder*, calls the library routines in order to print and save the analysis results. Separately, the program *LABELIT* (Sauter *et al.*, 2004) provides Python bindings to *DISTL*, allowing candidate Bragg spots to be used for autoindexing. *DISTL* accepts the image data as a matrix of pixel values, along with experimental information such as incident wavelength, incident beam position, detector distance and pixel size. The actual task of retrieving header information and pixel data from images of various formats is left to the users of the library, *e.g.* *Spotfinder* or *LABELIT*, which call appropriate libraries for reading X-ray data.

On a series of test images, the automated analysis provided a reliable alternative to visual inspection and was also consistent with a more rigorous integration of the Bragg spots using a conventional data-processing program. This frees the experimenters from the routine of inspecting every image so that they can focus on a small number of images that the program reports as being exceptionally good, bad, or otherwise deserving special attention.

**Table 1**
Main processing parameters and their default values.

| Symbol | Description | Default value |
|---|---|---|
| $n$ | Scan window edge length in pixels (square windows) | Three scans: 101, 51, 51 |
| – | Minimum proportion of background pixels in scan window | 2/3 |
| $\gamma_l$ | Background pixel $I$ (signal height) threshold | Three scans: 1.5, 2.0, 2.5 |
| $\gamma_u$ | Bragg spot pixel $I$ threshold | 3.8 |
| $\tau_1$ | Ice-ring pixel $I$ threshold value 1 | 0 |
| $\alpha_1$ | Ice-ring pixel $I$ fraction 1 | 0.55 |
| $\tau_2$ | Ice-ring pixel $I$ threshold value 2 | 1.5 |
| $\alpha_2$ | Ice-ring pixel $I$ fraction 2 | 0.20 |
| – | Minimum spot area in pixels (inclusive) | 5 |
| $\mu$ | Cutoff fraction in resolution estimation, method 2 | 15% |

## 2. Description of the program

The primary goal of the program is to evaluate the quality of a diffraction pattern. To this end, the program executes several steps in turn, addressing challenges posed by the intrinsic complexities of any diffraction image as well as by irregularities in specific crystals. Firstly, the Bragg scattering signal is extracted from the background noise in the diffraction pattern, taking into account the variation in local background resulting from both low-angle scattering and thermal diffuse scattering. The ultimate goal of this step is to identify the Bragg spots in order that they can be analyzed reliably. Secondly, the identified diffraction spots are validated to ensure that they are not the result of scattering from ice within the sample, which often manifests itself as a powder diffraction pattern of varying strength and granularity (ice-rings). Thirdly, for the purpose of automated sample evaluation and selection, the quality of each spot is gauged by its size and shape, and the entire image is evaluated in terms of the number, quality and distribution of diffraction spots, the presence of ice-rings or salt particles, and the limiting resolution of the diffraction pattern, which is estimated in a uniform fashion.

These steps are described below in detail. Table 1 lists the main processing parameters. The parameters are largely determined empirically and can be adjusted, although the default parameters work well in the majority of cases.

### 2.1. Calculating localized pixel signal heights

Each pixel on the image has a digitized value, $X$, which indicates the strength of diffraction at that location. Since the background varies with the scattering angle, the value $X$ is transformed into a measure of signal height, $I$, which symbolizes the significance of a specific pixel as compared to pixels in its neighborhood.

The neighborhood is taken as a square window of $n$ by $n$ pixels, centered on the pixel in question. Let $m_{n\times n}$ and $s_{n\times n}$ be the mean and standard deviation of the $X$ values of the neighborhood pixels, then $I$ is defined simply as the standardized $X$ value:

$$I(X) = \frac{X - m_{n\times n}}{s_{n\times n}}. \quad (1)$$

If $I(X)$ is below the threshold value $\gamma_l$, the pixel is regarded as background. Otherwise, the pixel potentially belongs to a diffraction spot.

The image is scanned three times; each scan provides a refinement based on the preceding one. In the first scan, all the pixels in the window contribute to the calculation of $m_{n\times n}$ and $s_{n\times n}$. On subsequent scans, only background pixels, as classified in the preceding scan, contribute. By restricting the calculation of local mean and standard deviation to background pixels, we avoid undue influence of large $X$ values on these statistics.

To ensure that $m_{n\times n}$ and $s_{n\times n}$ are representative of the background, background pixels must reach a pre-specified minimum fraction of the window (the default value of this fraction is 2/3; see Table 1); any window that does not meet this requirement is expanded until it does.

### 2.2. Detecting ice-rings

The types of ice-rings that *DISTL* currently handles are characterized by a concentration of strong pixels that, in the case of a flat orthogonal detector, reside on a circumference around the beam center. To detect such ice-rings, the image is divided into a series of thin circular shells centered on the direct-beam position. For a shell to be labeled as an ice-ring, the signal heights, $I$, of the pixels within the shell must satisfy the following conditions.

(i) For at least fraction $\alpha_1$ of the pixels, $I > \tau_1$.

(ii) For at least fraction $\alpha_2$ of the pixels, $I > \tau_2$, where $0 < \alpha_2 < \alpha_1 < 1$, and $\tau_1 < \tau_2$.

The rationale behind these requirements is that on the one hand, many (fraction $\alpha_1$) pixels in the shell are strong ($I > \tau_1$), and on the other hand, a fair fraction ($\alpha_2$, which is smaller than $\alpha_1$) are very strong ($I > \tau_2$, where $\tau_2 > \tau_1$) so that this shell would interfere with analysis of the image and has to be left out. Values of these parameters are chosen empirically. In terms of the default parameter values, a shell is labeled an ice-ring if at least 55% of its pixels have signal heights 0 or above, and at least 20% have signal heights 1.5 or above. If several contiguous shells all meet these requirements, they collectively represent one ice-ring. Since this approach is based on percentages of pixels, ice-rings in the corners of a square detector are handled in the same way as those in the region of full circular shells.

### 2.3. Identifying Bragg spots

Bragg spots are identified as patches of connected pixels that meet the following requirements.

(i) The signal height of each pixel in the patch is above the threshold $\gamma_u$.

(ii) The number of pixels in the patch is above a threshold spot size.

(iii) The patch lies entirely off any ice-ring.

By 'connected' we mean two pixels either share a border (assuming each pixel is a square) or are linked through pixels that share borders.

We shall call a pixel whose $X$ value is equal to or greater than the value of all its eight neighbor pixels a 'local maximum'. Each diffraction spot has at least one local maximum. Local maxima serve as the starting points in our spot searching algorithm. A spot also has a 'peak', *i.e.* the pixel with the largest $X$ value in the spot, and a 'center', *i.e.* the center of gravity weighted by pixel values. The number of local maxima and the location of the peak in a spot convey information on certain aspects of the diffraction quality of that spot.

### 2.4. Evaluating diffraction quality

At this point, the program has largely finished the task of identifying 'significant signals' on the image: candidate Bragg spots and ice-rings. Detailed information for each of these entities has been obtained. To facilitate automatic sample screening, the program calculates summary statistics that are indicative of the overall quality of the crystal sample. These statistics concern quality of the Bragg spots, severity of defects such as ice and salt in the crystal, and the limiting resolution of the diffraction pattern.

**2.4.1. Spot quality**. Besides size (*i.e.* pixel count) and peak height, three additional aspects of spot quality, namely shape, number of maxima and presence of close neighbors, are examined.

Ideally, we would like to use the spot shape as an indicator of crystal quality. Elliptical spot shapes can be produced by a number of factors of interest, such as intrinsic disorder in the crystal lattice, a split crystal, or even powder diffraction artifacts. In practice, however, an elliptical spot shape can also arise from factors unrelated to the crystal, such as anisotropic beam shape and divergence. The program makes no attempt to distinguish among these effects. Rather, it is intended that many crystals will be compared based on data collected under near-constant beam conditions. Therefore, when selecting the best crystals from a given batch, the ones with rounder spots will win. *DISTL* calculates a shape measure based on the coefficient of variation, CV (*i.e.* standard deviation divided by mean), of the distances from the border pixels of a spot to the center of the spot. The smaller the CV value, the more circular the spot.

On an ideal diffraction pattern most spots have one and only one local maximum. If a significant fraction of spots have multiple maxima, then either the crystal is cracked or the detector is too close to resolve the diffraction spots resulting from the crystal lattice. The program does not attempt to distinguish between these two causes. It simply reports the percentage of spots with multiple maxima.

The program considers two spots as 'close' if their peak-to-peak distance is less than 1.2 times the diameter of the larger spot of the two. The presence of close neighbors may suggest the same problems that cause multiple maxima.

**2.4.2. Strength of ice-rings**. Strong ice-rings can overshadow the diffraction in some resolution shells and hence significantly compromise a diffraction experiment. Given several samples with comparable resolution, it is usually preferable to select the sample with the least ice contamination.

Suppose signal heights of the $n$ pixels on an ice-ring are $\{I_i\}$, where $i = 1, \ldots, n$. Let $p_1 = \#\{I_i\colon I_i > \tau_1\}$ and $p_2 = \#\{I_i\colon I_i > \tau_2\}$, where $\#\{\ldots\}$ stands for 'the number of...'. The program defines the strength of the ice-ring as

$$0.6\,\frac{p_1/n - \alpha_1}{1 - \alpha_1} + 0.4\,\frac{p_2/n - \alpha_2}{1 - \alpha_2}. \qquad (2)$$

Parameters $\tau_1$, $\tau_2$, and $\alpha_1$, $\alpha_2$ are defined in §2.2. Expression (2) measures how well the two ice-ring criteria (§2.2) are exceeded. As with the ice-ring detection method, ice-rings in the corners of a square detector are not special cases with the measure above.

**2.4.3. Overloaded patches**. Sometimes a salt crystal rather than the desired macromolecular sample is mistakenly frozen. This is common for certain crystallization conditions. Diffraction patterns from salt are typically characterized by a few extremely strong diffraction peaks, which often exceed the regular range of image pixel intensity and are represented by overloaded patches on the diffraction image.

Overloaded patches can also result from an inaccurately positioned X-ray beam-stop, which in normal operation prevents the direct beam from hitting the detector. An exposure time that is too long may also cause overloaded patches.

The program reports size and location of overloaded patches, if any, to indicate possible existence of such problems.

**2.4.4. Diffraction resolution**. In the context of discussing published crystal structures, limiting resolution conventionally refers to a property of a complete data set, not a single image. Before it can be calculated, integrated Bragg spot intensities must be obtained, Lorentz and polarization corrections must be applied, partial reflections must be scaled or summed, and multiple measurements must be merged. In contrast, *DISTL* addresses specific situations where a complete data set has not been (or cannot be) processed, including (*a*) selecting the best crystals from a batch, based on one or two oscillation images collected from each sample; (*b*) monitoring the diffraction quality of a data set in real time, as each successive oscillation image is acquired; and (*c*) choosing candidate Bragg spots for autoindexing. To support these goals, it is useful to assess the limiting resolution of a single unprocessed image, just as a practicing crystallographer does intuitively when inspecting the data visually.

In considering our working definition of image resolution, we decided to limit ourselves to an isotropic model. This is only an approximation because the Lorentz correction has not yet been applied, so Bragg spots near the spindle will be overrepresented. (Also, the diffraction pattern may be truly anisotropic.) However, this approximation is generally sufficient for comparing statistics between crystal samples.

Since integrated intensities for all diffraction spots are not available at this stage, we examine the number of spots and their distribution instead. In the diffraction image of a perfect crystal, the spot density $f$ has a known, albeit complicated, functional dependence on the resolution $D$ and the oscillation range. The dependency ranges from $\mathrm{d}f \propto D^{-2}$ for a still photograph, to $\mathrm{d}f \propto D^{-3}$ for a Laue photograph or 360° oscillation shot, where $\mathrm{d}f$ is the number of spots lying between resolutions $D$ and $D + \mathrm{d}D$. The variable $\mathrm{d}f$ is also expected to be a function of the effective mosaic spread.

In practice we do not have perfect crystals and we do not rigorously analyze the observed spot density function. Rather we aim to calculate a reproducible quantity consistent with conventional expectations.

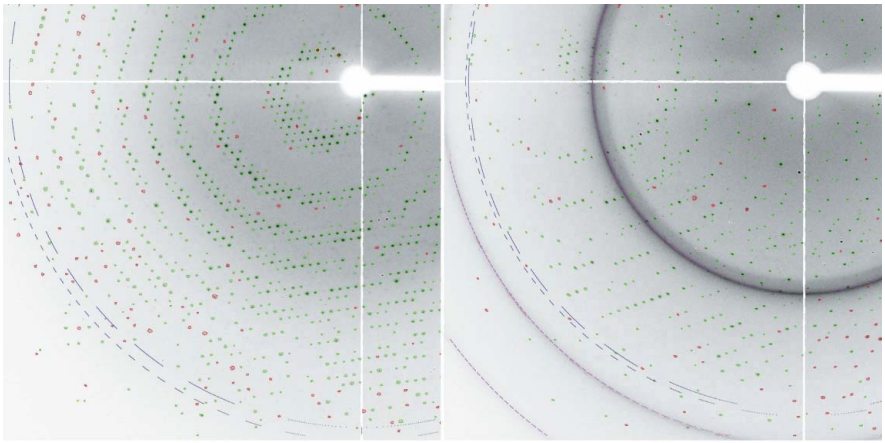
The program implements two methods, both loosely based on the linear relationship assumption $\mathrm{d}f \propto D^{-3}$. Method 1 orders the spots by resolution, selects spots at constant intervals of spot number, and examines the resolution of the selected spots. Method 2 evenly divides the reciprocal space and examines the trend of the number of spots in each space interval. Both methods use a list of 'good' spots that exclude spots with overloaded pixels and spots whose diffraction angle $2\theta$ is smaller than 2.9°. The second exclusion is meant to avoid interference by noise near the beam center.

**Method 1**. All spots are ordered by scattering angle and from this list an equi-spaced (in terms of number of spots) subset of $n$ spots are sampled. Ideally, the series $D_i^{-3}$, where $i = 1, \ldots, n$, of these sampled spots should increase linearly, as the ideal model of equal numbers of spots in constant $D^{-3}$ intervals would suggest. In reality, the series forms a concave (bent upward) curve because fewer spots are detectable at higher resolutions, as weak spots become less distinguishable from noise and the background.

The program takes the 'elbow', *i.e.* the point where the curve starts to diverge significantly from its expected linear trend, as an estimate of the diffraction resolution of the image. Finding the elbow of a curve has been the subject of some research efforts. We borrow ideas from Tibshirani's *Gap statistic* (Tibshirani *et al.*, 2001). The algorithm proceeds as follows.

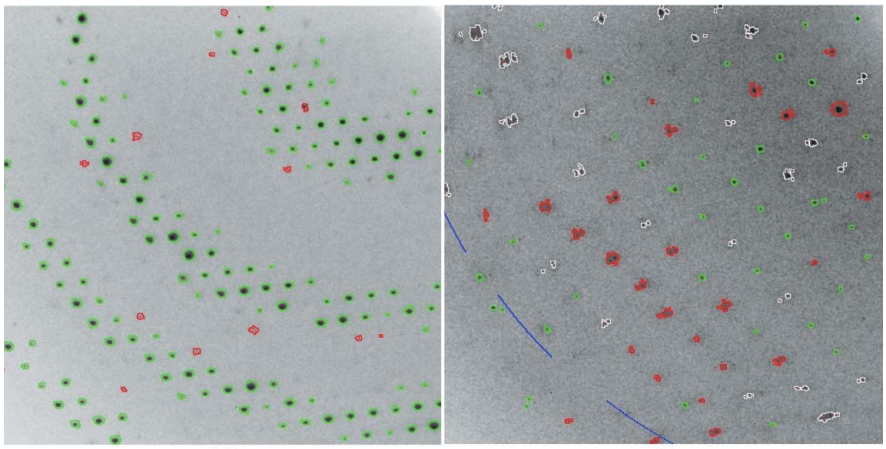
(i) On the Cartesian plane, plot points $P_i = (i, D_i^{-3})$, where $i = 1, \ldots, n$ is the serial number of spots in the subset. Find the point, say $P_m$ where $m \geq 2$, whose connection with point $P_1$ has the largest slope. Points $P_1, P_2, \ldots, P_m$ along with the straight line $P_1P_m$ are shaped like an archer's bow (see Fig. 3). It is then assumed that the image resolution is between $D_1$ and $D_m$. Usually, $P_m$ is the last point, that is, $m = n$. But $m$ could be smaller than $n$ if an unusually high number of spots are detected around a certain intermediate resolution. In that case, our search for the image resolution does not go outside the spot 'bump'. This is particularly useful when ice-rings are present.

(ii) Calculate the vertical distances $G_i$, called 'gaps', from points $P_i$ to the straight line $P_1P_m$, where $i = 2, \ldots, m - 1$. Then calculate the standard deviation $s$ of the gaps.



(*a*) (*b*)

**Figure 1**
Two processed example images. Types of spots are indicated by border color: green represents good; red has multiple maxima; yellow indicates an overloaded peak; white has close neighbors. Yellow pixels are overloaded. Red pixels are local maxima. Dashed magenta lines are ice- rings (the dashed line for the thick ice-ring may look solid). The blue line with short dashes is the resolution limit by method 1. The blue line with long dashes is the resolution limit by method 2.



(*a*) (*b*)

**Figure 2**
(*a*) Details of Fig. 1(*a*). (*b*) Details of an image with bad spots. Spots with white borders have close neighbors. Spots with red borders have multiple maxima.

(iii) Find the largest gap, say $G_k$ at point $P_k$. Then identify points $P_i$ to the right of $P_k$ (*i.e.* $i > k$) that satisfy $G_i > G_k - 0.5s$. Take the far-right one of these points, say $P_g$, as the elbow point, and its resolution $D_g$ as an estimate of the image resolution. In essence, this step finds the point with the largest gap and takes a point to the right of it whose gap is still quite large – within half a standard deviation below the largest.

**Method 2**. Recall that *DISTL* identifies spots using a signal height threshold ($\gamma_u$ in Table 1). Therefore, counting the number of spots in different resolution shells implicitly corresponds to analyzing the signal heights in a resolution shell relative to a fixed $\sigma$ cutoff. If the number of spots per resolution shell falls below some expected value, this suggests the limiting resolution.

In this method, the reciprocal space is divided into resolution shells of equal reciprocal-space volume. The shell volume is chosen such that the lowest resolution shell (closest to the beam center) contains 5% of the spots, or 25, whichever is greater. When the spot counts $t_i$, $i = 1, \ldots, m$, for $m$ shells are plotted, we often notice that $t_1$, the spot count in the lowest resolution shell, is the highest and that a plateau value is reached later. We use $t_0 = (t_1 + t_2)/2$ as a reference and define the limiting resolution as the outer boundary resolution of shell $j$, where $j$ is the smallest index such that $t_j < \mu t_0$ and $t_{j+1} < \mu t_0$ (see Fig. 3). In other words, we find the first two consecutive shells whose spot numbers both fall below $\mu t_0$, and use the first of these two shells to define the image resolution. The default value for the cutoff percentage $\mu$ is 15%.

**Other considerations**. Because of the rectangular geometry of the detector, spots at high resolution are not all recorded on the image; when part of the circle corresponding to a high resolution falls out of the image, spots at that resolution are available in the corners only. In both methods we correct for this artifact by inserting an appropriate number of dummy spots at certain locations of the resolution-ordered spot list. Consequently, if an image has high-quality diffraction spots throughout, the estimated image resolution correctly extends into the corners.

In most cases method 1 and method 2 give comparable estimates. If a diffraction image is notably weak or irregular, two situations can arise. In the first situation, both estimates are very low or even unobtainable. Although the low resolution estimates should not be taken literally, the user can be confident that the image is noisy. In the second situation, the two estimates differ greatly, suggesting that neither is reliable. This prompts us to provide a measure of reliability for the estimates. In method 1, the slopes $k_i$ of the lines $P_1P_i$, where $i = 2, \ldots, n$ [$n$ as used in the description of step (i) in method 1], are expected to increase monotonically on a reasonable image. We measure the 'noisiness' of this series by

$$\frac{\#\{(k_i, k_j)\colon 2 \le i < j \le n \text{ and } k_i \ge k_j\}}{(n-1)(n-2)/2}, \qquad (3)$$

where #{ ... } means 'the number of ...'. Noticing that $(n-1)(n-2)/2$ is the total number of pair-wise comparisons between the $k_i$'s, the expression above is the fraction of 'overturned' slope pairs. Similarly, for method 2, the noisiness is measured by

$$\frac{\#\{(t_i, t_j)\colon 1 \le i < j \le m \text{ and } t_i \le t_j\}}{m(m-1)/2}. \qquad (4)$$

These noisiness measures hint at not only the reliability of the resolution estimates, but also the overall quality of the image.

**2.4.5. Program output**. *DISTL* is designed as a class library that processes a diffraction image and passes image information on to the caller program. It provides the following information about the image.

(i) Original value ($X$), signal height ($I$), and local background value ($m_{n\times n}$) of each pixel.

(ii) A list of spots. Information for each spot includes lists of its internal pixels, border pixels, and local maxima; location of the spot's peak and center; resolution at the spot peak; measure of shape; and number of close neighbors.

**Table 2**
Processing results for the example images in Fig. 1.

| Description | Fig. 1(*a*) | Fig. 1(*b*) |
|---|---|---|
| Number of spots | 2092 | 1133 |
| Number of spots with overloaded pixels | 6 | 0 |
| Number of spots with close neighbors | 4 | 15 |
| Number of spots with multiple maxima | 43 | 17 |
| Median spot area in pixel counts | 22 | 16 |
| Median spot shape | 0.87 | 0.71 |
| Ice-rings (Å) | None | 3.66–3.62, 2.24–2.22, 1.91–1.90 |
| Strength of the strongest ice-ring | N/A | 0.83 |
| Size of the largest overloaded patch | 4 | None |
| Resolution limit by method 1 (Å) | 1.95 | 2.65 |
| Resolution limit by method 2 (Å) | 2.02 | 2.36 |
| Noisiness of the curve in method 1 | 0.004 | 0.024 |
| Noisiness of the curve in method 2 | 0.067 | 0.085 |

(iii) A list of local maxima. Maxima are of type 'point' (or pixel), containing information on location, value ($X$) and signal height ($I$).

(iv) A list of overloaded patches. Patches are of type 'spot' and contain the same information as spots. For an overloaded patch we are mainly interested in its area, location, and whether or not it falls on an ice-ring.

(v) A list of ice-rings. Information about an ice-ring includes its resolution boundaries (corresponding to both edges of the shell), measure of strength and number of pixels.

(vi) Estimated limiting resolution. Estimates and reliability measures (noisiness) with both method 1 and method 2 are provided, along with the series $D_i$, $i = 1, \ldots, n$, for method 1, and $t_i$, $i = 1, \ldots, m$, for method 2.

The caller program may choose to combine the above information into an overall quality score for the diffraction image.

The light-weight program *Spotfinder* works closely with the library and outputs a series of statistics and processed images. Output of primary interest includes the following.

(*a*) A processed image with spots, local maxima, ice-rings and limiting resolution marked out. See Fig. 1 for two examples.

(*b*) A log file containing values of processing parameters and a variety of statistics of the image, including the number of spots by categories (all spots, overloaded spots, spots with close neighbors, spots with multiple maxima), median of spot size, median of spot shape, number of local maxima, size of the largest overloaded patch (and whether or not it is on an ice-ring), location (*i.e.* boundary resolutions) of ice-rings, strength of the strongest ice-ring, estimated limiting resolution (*via* methods 1 and 2) and reliability of the estimates.

## 3. Validation of the program

The program was developed using a set of diffraction images collected from 50 different crystals of proteins produced by the Joint Center for Structural Genomics (Lesley *et al.*, 2002). The images were collected using Quantum 4 and Quantum 315 detectors on beamlines 9–2 and 11–1 at the Stanford Synchrotron Radiation Laboratory.

To demonstrate some of the key output, we show the processing results of one strong image (Fig. 1*a*) and one weak image (Fig. 1*b*). The program detects three ice-rings on the weak image and marks them with magenta lines. Also shown are the resolution limits by method 1 and method 2. Notice that strong pixels near the ice-rings are not regarded as spots. Main processing results for these images are listed in Table 2. Fig. 2 zooms in on Fig. 1(*a*) and on another image where many spots are bad, either with close neighbors or with multiple maxima.

Fig. 3 shows the $(i, D_i^{-3})$ and $(i, t_i)$ curves used by methods 1 and 2 for estimating resolution limits. Location of the estimated resolution is indicated by a vertical dashed line.

We tested the program on an independent set of images to ensure that the ice-ring and resolution-limit determination was consistent with results provided by an experienced human crystallographer. In addition, we wanted to determine how the resolution limits from methods 1 and 2 compare with the more traditional procedure of calculating resolution limit using integrated Bragg spot intensities. Although methods 1 and 2 are intended to be used only in situations where indexing and integration have not yet been performed, it would be most convenient if the limiting resolutions estimated before and after integration are correlated.

Diffraction images from cryo-cooled samples were acquired at ALS (Advanced Light Source) beamline 5.0.3 using 1.0 Å radiation and an ADSC (Area Detector Systems Corporation) Quantum 4R detector at a crystal-to-detector distance of 250 mm. Exposure times varied from 20 to 90 s for 1° oscillation photographs. The samples represented six different crystal types. Visual inspection was performed by a single individual using the program *ADXV* (from ADSC), and images were graded as to their estimated limiting resolution, ice-ring content, spot shape and diffraction strength. Bragg spots from §2.3 were used to autoindex the images (Sauter *et al.*, 2004), and reflections were integrated with the program *MOSFLM* (Leslie, 2001). For the purpose of producing the list of Bragg peaks for autoindexing, the resolution limit from method 2 was used as a cutoff. However, after autoindexing was complete, images were integrated out to the edge of the detector. Integrated intensities from fully and partially recorded Bragg reflections were combined into one list, and ranked according to resolution. The running average of the signal-to-noise ratio ($I/\sigma$, not to be confused with the same symbols used earlier) was computed with a window size of 4% of the total number of integrated reflections, or 20, whichever was greater. Limiting resolution was defined as that point where the running average fell below 1.5. This $\sigma$ cutoff was chosen so that the limiting resolution from visual inspection would roughly coincide with that from integrated intensities (Fig. 4). Altogether we obtained 276 images with resolutions high enough for autoindexing, but low enough so that the limiting resolution diffraction spots were properly collected with this detector geometry.

It is perhaps surprising that the assessment of limiting resolution by spot integration and visual inspection produces such a fuzzy correlation (Fig. 4). For example, diffraction patterns rated at 3.5 Å resolution by spot integration were judged to diffract anywhere from 2.8 to 4.0 Å by visual inspection. Part of this spread can be attributed to the subjectivity of the experimentalist (the original reason for seeking automated analysis methods). However, even the single set of results from spot integration gave different estimates when analyzed in different ways. When the limiting resolution was determined by dividing the integrated spots into resolution bins instead of using a running average, 3.5 Å images were rated anywhere from 3.3 to 3.8 Å (data not shown). This gives an idea of the limiting accuracy with which the resolution limit can be known.

These considerations must be taken into account when inspecting the resolution-limit results from methods 1 and 2. As seen in Fig. 5, the two methods yielded roughly correlated resolution estimates; correlated to the degree generally expected from our comparison of the two other methods in Fig. 4. Furthermore, both methods produced resolution-limit results that are usefully correlated with the more rigorous resolution estimate from integrated intensities (Fig. 6). One difference between the methods is that method 2 produces a systematically lower resolution estimate. Indeed, a primary motiva-

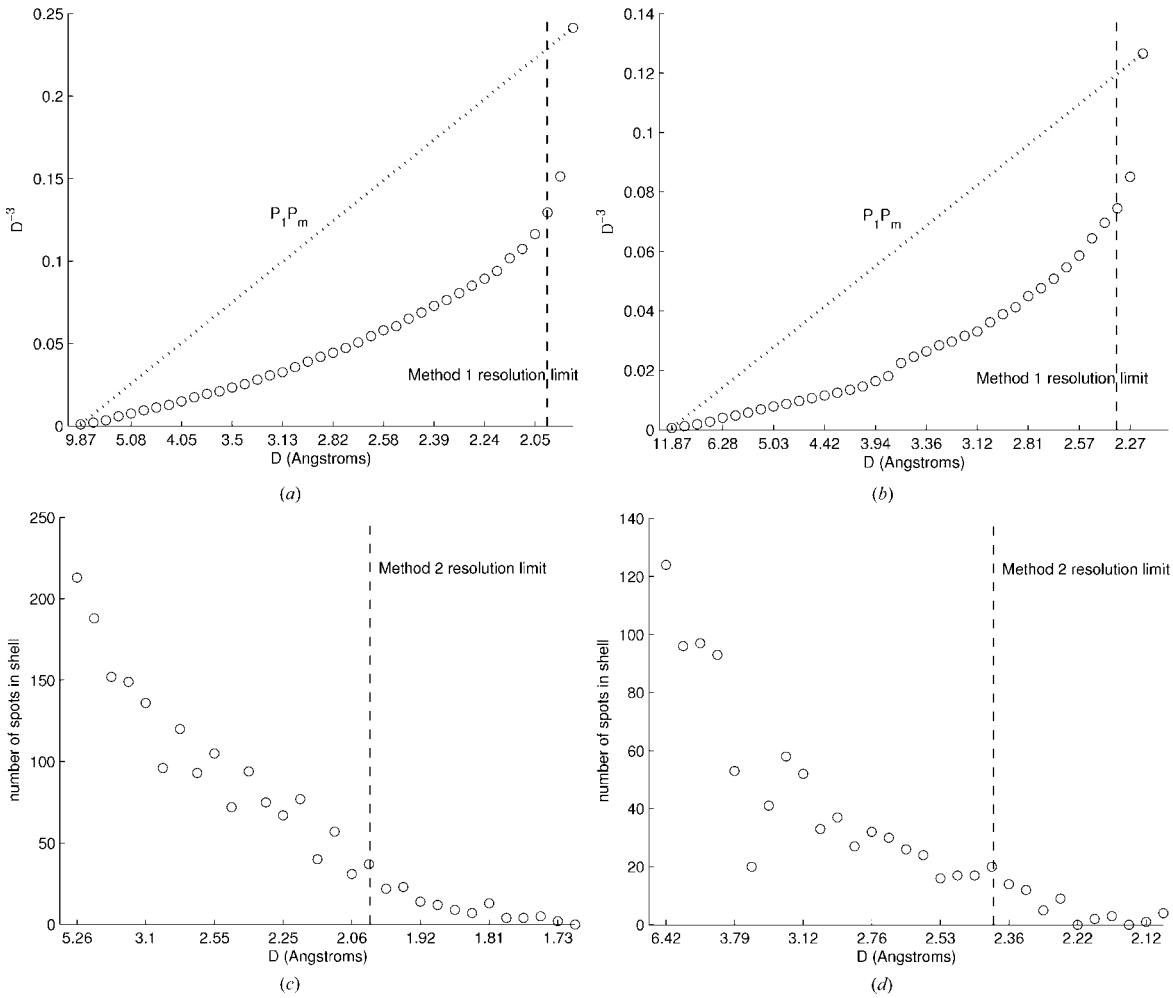



**Figure 3**
(*a*) The *i* versus $D_i^{-3}$ curve of the limiting-resolution estimation method 1 for the image in Fig. 1(*a*). (*b*) The *i* versus $D_i^{-3}$ curve of method 1 for the image in Fig. 1(*b*). (*c*) The *i* versus $t_i$ curve of method 2 for the image in Fig. 1(*a*). (*d*) The *i* versus $t_i$ curve of method 2 for the image in Fig. 1(*b*). The abscissa is the sequential index *i*, increasing from left to right, but is labeled by the corresponding resolution value, which decreases from left to right. At each index is a spot in method 1, or a resolution shell in method 2. See the text for step (i) of method 1 for explanations of the line $P_1P_m$.

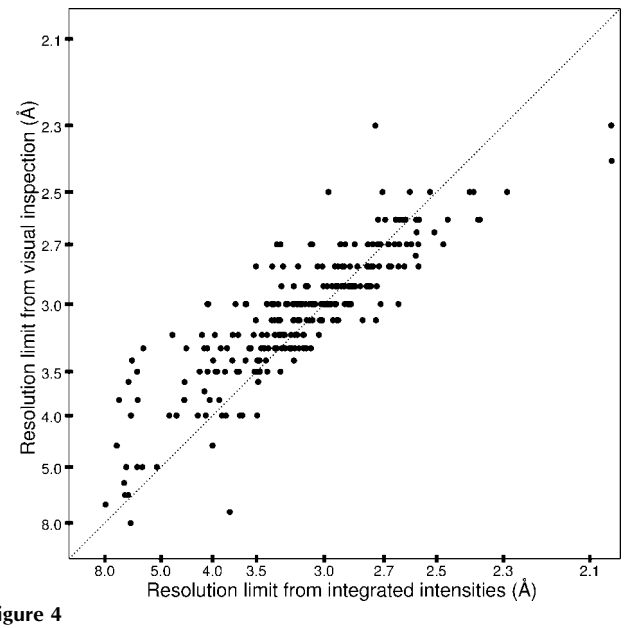



**Figure 4**
Comparison of resolution limits obtained by visual inspection and by analyzing integrated intensities.

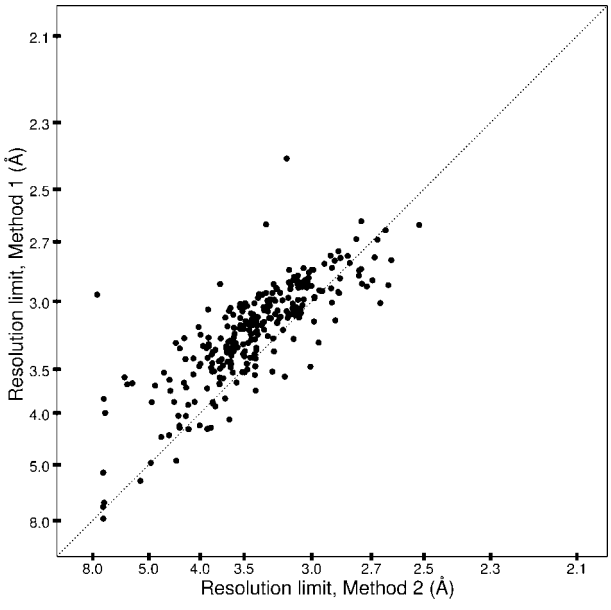



**Figure 5**
Comparison of resolution limits obtained by method 1 and method 2.

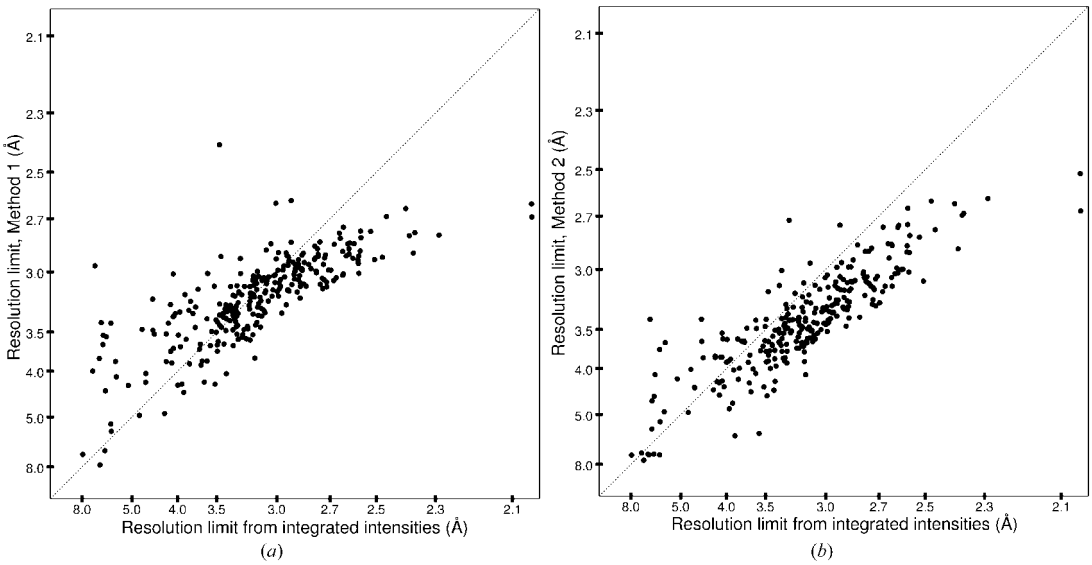



**Figure 6**
Comparison of resolution limits obtained by *DISTL* and by analyzing integrated intensities. (*a*) Method 1 in *DISTL*. (*b*) Method 2 in *DISTL*.

**Table 3**
Results of ice-ring detection on images.

| | | *DISTL* | |
|---|---|---|---|
| | | Yes | No |
| Visual inspection | Yes | 15 | 4 |
| | No | 3 | 254 |

tion for developing method 2 was to produce a more conservative resolution cutoff for listing candidate Bragg spots for autoindexing. Indexing can fail in a small percentage of cases if the cutoff resolution is too high.

Positive results were also obtained for *DISTL*'s ice-ring detection. As shown in Table 3, *DISTL* agreed with visual inspection that 269 images either did or did not contain ice-rings. In the seven instances where there was disagreement, further inspection forced us to conclude that *DISTL*'s interpretation was preferred.

## 4. Conclusions

Rapid analysis of oscillation photographs will play an important role in the present and future automation of macromolecular crystallography experiments (Criswell *et al.*, 2004; Sauter *et al.*, 2004). The *DISTL* package provides a quick summary of the X-ray diffraction image, including a list of candidate Bragg spots, statistical properties such as average spot intensity, spot shape and limiting resolution, as well as an assessment of pathologies including powder patterns and pixel overloads. Typical response time for a 10 Mbyte image is 2.5 s on a 2.8 GHz Intel processor running Linux, which is quick enough for the analysis results to be used for experimental decisions in real time.

At the Stanford Synchrotron Radiation Laboratory (SSRL), *Spotfinder* has been incorporated into the data management program *Web-Ice* (González *et al.*, 2005), allowing users to browse interactively the image quality statistics calculated by *DISTL*. The *Web-Ice* system also includes the program *LABELIT* (Sauter *et al.*, 2004), which is configured to accept Bragg spots from *DISTL* for autoindexing. Indexing performed with this set of candidate spots is very robust, since *DISTL* eliminates several types of artifacts (including ice-rings and split spots) and sets a useful cutoff for the resolution limit. The indexing solution of *LABELIT* is then used as a basis for integration with *MOSFLM* (Leslie, 2001). All of these results can be promptly viewed within the *Web-Ice* display. Plans exist to link this information with other programs for determining data collection geometry (Ravelli *et al.*, 1997), and for setting proper data collection parameters so that data with a given level of experimental significance can be acquired (Popov & Bourenkov, 2003). A similar system will also be implemented at the Advanced Light Source.

In the future, *DISTL* and *Spotfinder* will be expanded to be more widely applicable, *e.g.* to images acquired by a detector swung out to a non-zero $2\theta$ angle. The software may be obtained by contacting the authors.

The authors thank all members of the JCSG Structure Determination Core at SSRL for providing test data. In particular, we gratefully acknowledge Qingping Xu and Günter Wolf for their feedback and suggestions. Also, we thank Ana González, Penjit Moorhead and Scott McPhillips for their collaboration involving *Web-Ice*. The JCSG is funded by the Protein Structure Initiative of the National Institute of Health and National Institute of General Medical Sciences (grant P50 GM62411). SSRL operations are funded by DOE BES and the SSRL Structural Molecular Biology program by DOE BER, NIH NCRR BTP and NIH NIGMS. NKS acknowledges support from NIH/NIGMS under grant number 1P50GM62412. Lawrence Berkeley National Laboratory is funded in part by the US Department of Energy under Contract No. DE-AC03-76SF00098.

## References

Cohen, A. E., Ellis, P. J., Miller, M. D., Deacon, A. M. & Phizackerley, R. P. (2002). *J. Appl. Cryst.* **35**, 720–726.
Criswell, A. R., Bolotovsky, R., Niemeyer, T., Athay, R. & Pflugrath, J. W. (2004). *Acta Cryst.* A**60**, s112.
González, A., Moorhead, P., McPhillips, S. & Sauter, N. K. (2005). *Acta Cryst.* A**61**, C486.
Lesley, S. A., Kuhn, P., Godzik, A., Deacon, A. M., Mathews, I., Kreusch, A., Spraggon, G., Klock, H. E., McMullan, D., Shin, T., Vincent, J., Robb, A., Brinen, L. S., Miller, M. D., McPhillips, T. M., Miller, M. A., Scheibe, D.,

Canaves, J. M., Guda, C., Jaroszewski, L., Selby, T. L., Elsliger, M.-A., Wooley, J., Taylor, S. S., Hodgson, K. O., Wilson, I. A., Schultz, P. G. & Stevens, R. C. (2002). *PNAS*, **99**, 11664–11669.
Leslie, A. G. W. (2001). *International Tables for Crystallography: Volume F, Crystallography of Biological Macromolecules*, edited by M. G. Rossmann & E. Arnold, pp. 212–217. Dordrecht: Kluwer Academic Publishers.
McPhillips, T. M., McPhillips, S. E., Chiu, H.-J., Cohen, A. E., Deacon, A. M., Ellis, P. J., Garmen, E., González, A., Sauter, N. K., Phizackerley, R. P., Soltis, S. M. & Kuhn, P. (2002). *J. Synchrotron Rad.* **9**, 401–406.
Miller, M. D., Brinen, L. S., Cohen, A., Deacon, A. M., Ellis, P., McPhillips, S. E., McPhillips, T. M., Phizackerley, R. P., Soltis, S. M., van den Bedem, H., Wolf, G., Xu, Q. & Zhang, Z. (2004). *AIP Conf. Proc.* **705**, 1233–1236.
Muchmore, S. W., Olson, J., Jones, R., Pan, J., Blum, M., Greer, J., Merrick, S. M., Magdalinos, P. & Nienaber, V. L. (2000). *Structure*, **8**, R243–R246.
Popov, A. N. & Bourenkov, G. P. (2003). *Acta Cryst.* D**59**, 1145–1153.
Ravelli, R. B. G., Sweet, R. M., Skinner, J. M., Duisenberg, A. J. M. & Kroon, J. (1997). *J. Appl. Cryst.* **30**, 551–554.
Sauter, N. K., Grosse-Kunstleve, R. W. & Adams, P. D. (2004). *J. Appl. Cryst.* **37**, 399–409.
Snell, G., Cork, C., Nordmeyer, R., Cornell, E., Meigs, G., Yegian, D., Jaklevic, J., Jin, J. & Earnest, T. (2004). *Structure*, **12**, 537–545.
Stevens, R. C., Yokoyama, S. & Wilson, I. A. (2001). *Science*, **294**, 89–92.
Tibshirani, R., Walther, G. & Hastie, T. (2001). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **63**, 411–423.