Acta Cryst. (1989). A**45**, 573-576

R Factors in X-ray Fiber Diffraction. II. Largest Likely *R* Factors

By R. P. Millane

The Whistler Center for Carbohydrate Research, Smith Hall, Purdue University,
West Lafayette, Indiana 47907, *USA*

(*Received* 4 *January* 1989; *accepted* 21 *March* 1989)

Abstract

The largest likely *R* factor (that for a structure uncorrelated with the correct structure) is smaller for an X-ray fiber diffraction analysis than for a traditional single-crystal analysis. For example, the largest likely *R* factor for tobacco mosaic virus determined by fiber diffraction at 3 Å resolution is 0·31, compared to 0·59 for a single-crystal analysis. Earlier treatments of largest likely *R* factors in fiber diffraction for a fixed number of overlapping Fourier–Bessel structure factors are extended to general fiber diffraction patterns. The theory is illustrated with applications to particular structures thereby elucidating some general features of fiber diffraction *R* factors. These results are useful for interpreting the reliability of structure determinations, and may also be useful for further developments of fiber diffraction theory in general.

1. Introduction

The *R* factor is a useful measure of the quality of structures determined by both traditional crystallography and fiber diffraction. The significance of the *R* factor obtained for a particular structure can be assessed by comparing it with the largest likely *R* factor; that for a structure uncorrelated with the correct structure. The largest likely *R* factor for single crystals was determined by Wilson (1950) and has recently been examined for fiber diffraction (Stubbs, 1989; Millane, 1989). In fiber diffraction, intensity measurements are sums of the intensities of a number of Fourier–Bessel structure factors (Klug, Crick & Wyckoff, 1958), the number varying over the diffraction pattern and depending on the diameter and symmetry of the diffracting particle. The largest likely *R* factor therefore depends on the molecular diameter and symmetry, and the maximum resolution on the diffraction pattern. It is convenient to consider its calculation in two parts: calculation of largest likely *R* factors for fixed numbers of Fourier–Bessel terms, followed by the use of these to calculate the largest likely *R* factor for a particular diffraction pattern. Largest likely *R* factors for fixed numbers of terms have been determined by Stubbs (1989) and Millane (1989), but the second part has been defined only approximately (Stubbs, 1989). The relationship between the *R* factor for a fiber diffraction pattern and *R* factors for fixed numbers of overlapping terms is derived here. This allows accurate calculation of largest likely *R* factors for general fiber diffraction patterns.

An expression for the largest likely *R* factor in fiber diffraction is developed in the next section. In the following section, the theory is illustrated by applications to particular structures and the results discussed.

2. Theory

The *R* factor is given by

$$R=\sum_{i=1}^{N}|F_i-F_i^o|\bigg/\sum_{i=1}^{N}F_i^o=\langle|F-F^o|\rangle/\langle F^o\rangle \quad (1)$$

where F_i and F_i^o are the calculated and observed structure amplitudes respectively, $\langle\ \rangle$ represents ensemble averaging in reciprocal space and there are N measurements. In traditional crystallography, F_i and F_i^o are the individual structure amplitudes. In fiber diffraction, however, because the diffracting particles or crystallites are randomly rotated, the diffraction pattern is cylindrically averaged. The measured amplitudes are therefore equal to the square roots of the sums of a number of intensities. For a non-crystalline specimen, the measurements are samples (along layer lines) of the cylindrically averaged continuous transform of the diffracting particle. The number m of independent terms averaged depends on the diameter and symmetry of the diffracting particle, and the position in reciprocal space. It is convenient to denote the measured amplitude by the length $\mathcal{G}$ of an m-dimensional vector $\boldsymbol{\mathcal{G}}$ whose components are the real and imaginary parts of the complex Fourier–Bessel structure factors G_n (Stubbs, 1989; Millane, 1989). For a polycrystalline specimen, each measurement is a set of composite crystalline intensities, the number depending on the space group, the cell constants, the mean crystallite size and disorientation (since these affect the overlap of adjacent reflections), and the position in reciprocal space, and $\mathcal{G}$ is used to represent the amplitude of a composite reflection.

In fiber diffraction, therefore, F in (1) is replaced by $\mathcal{G}$. Defining $\Delta\mathcal{G}=|\mathcal{G}-\mathcal{G}^o|$, noting that $\langle\mathcal{G}^o\rangle=\langle\mathcal{G}\rangle$, and grouping together amplitudes that have the same

Table 1. *Values of R_m, S_m and R_mS_m used in equation* (8) *to calculate largest likely R factors*

	R_m		S_m		R_mS_m
m	Exact	Approximate	Exact	Approximate	Approximate
1	$2\sqrt{2}-2$	0·828	$1/\sqrt{\pi}$	0·564	0·467
2	$2-\sqrt{2}$	0·586	$\sqrt{\pi}/2$	0·886	0·519
3	$7\sqrt{2}/4-2$	0·475	$2/\sqrt{\pi}$	1·128	0·536
4	$2-9\sqrt{2}/8$	0·409	$3\sqrt{\pi}/4$	1·329	0·544
5	$107\sqrt{2}/64-2$	0·364	$8/(3\sqrt{\pi})$	1·505	0·548
6	$2-151\sqrt{2}/128$	0·332	$15\sqrt{\pi}/16$	1·662	0·551
7	$835\sqrt{2}/512-2$	0·306	$16/(5\sqrt{\pi})$	1·805	0·553
8	$2-1241\sqrt{2}/1024$	0·286	$35\sqrt{\pi}/32$	1·939	0·555
9	$26291\sqrt{2}/16384-2$	0·269	$128/(35\sqrt{\pi})$	2·063	0·556
10	$2-40427\sqrt{2}/32768$	0·255	$315\sqrt{\pi}/256$	2·181	0·557
11		0·243		2·293	0·557
12		0·233		2·399	0·558
13		0·223		2·501	0·558
14		0·215		2·599	0·559
15		0·208		2·693	0·559
16		0·201		2·785	0·560
17		0·195		2·873	0·560
18		0·189		2·959	0·560
19		0·184		3·042	0·560
20		0·180		3·123	0·561

value of m, one may put (1) in the form

$$R = \langle N_m \langle \Delta\mathscr{G}\rangle_m\rangle / \langle N_m \langle \mathscr{G}\rangle_m\rangle \qquad (2)$$

where N_m denotes the number of data with m overlapping terms, $\langle\,\rangle_m$ represents averaging of the $\mathscr{G}$ that contain m terms, and $\langle\,\rangle$ represents averaging over m. If every amplitude contained m terms, then the R factor, denoted by R_m, would be given by

$$R_m = \langle \Delta\mathscr{G}\rangle_m / \langle \mathscr{G}\rangle_m. \qquad (3)$$

The cases $m=1$ and 2 in (3) correspond to centric and acentric single crystals respectively. For a random structure, largest likely values for R_m have been determined by Stubbs (1989) and Millane (1989) for any value of m. Since m is not constant on a fiber diffraction pattern, the R_m cannot be used directly. However, R factors for general fiber diffraction patterns can be calculated, using the R_m, as follows. Use of (2) and (3) shows that

$$R = \langle N_m R_m \langle \mathscr{G}\rangle_m\rangle / \langle N_m \langle \mathscr{G}\rangle_m\rangle \qquad (4)$$

or

$$R = \sum_{m=1}^{M} N_m R_m \langle \mathscr{G}\rangle_m \Big/ \sum_{m=1}^{M} N_m \langle \mathscr{G}\rangle_m \qquad (5)$$

where M is the maximum value of m on the diffraction pattern. Stubbs (1989) indicated that the R factor on a diffraction pattern can be estimated as a weighted (by N_m/N) average of the R_m, which gives an expression somewhat different from (5).

To calculate the largest likely value of R, largest likely values for the R_m are substituted into (5). These are given by (Millane, 1989)

$$R_m = 2 - 2^{m+2} m \binom{2m-1}{m} B_{1/2}(m/2+1/2, m/2) \qquad (6)$$

and [by the use of equations (7) and (11) of Millane (1989) and the gamma function $\Gamma(m)$]

$$\langle \mathscr{G}\rangle_m = \varepsilon^{1/2}\Gamma(m/2+1/2)/\Gamma(m/2) \qquad (7)$$

where

$$\binom{m}{n}$$

is the binomial coefficient, $B_x(m, n)$ is the incomplete beta function, and ε can be estimated from the atomic scattering factors (Stubbs, 1989). Inspection of (5) and (7) shows that R is independent of ε so that (5) can be written as

$$R = \sum_{m=1}^{M} N_m R_m S_m \Big/ \sum_{m=1}^{M} N_m S_m \qquad (8)$$

where

$$S_m = \varepsilon^{-1/2}\langle \mathscr{G}\rangle_m = \Gamma(m/2+1/2)/\Gamma(m/2). \qquad (9)$$

Using (6) and (9) and simplifying, one finds

$$\begin{aligned} R_mS_m = {} & [2/\Gamma(m/2)]\{\Gamma(m/2+1/2) \\ & -[2m\Gamma(m+1/2)/\Gamma(m/2) \\ & \times B_{1/2}(m/2+1/2, m/2)\}. \end{aligned} \qquad (10)$$

Values of R_m, S_m and R_mS_m are listed in Table 1 for

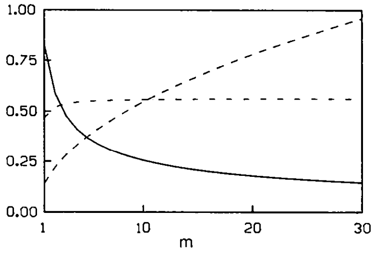

Fig. 1. Dependence of R_m (solid curve), $S_m/4$ (dashed curve) and R_mS_m (chain curve) on the number of overlapping terms m.

Table 2. *Largest likely R factors for four structures*

Molecule	Helix symmetry	Maximum radius (Å)	*c* repeat (Å)	Minimum resolution (Å)	Maximum resolution (Å)	*M*	*R*
K^+C-4-S	3_2	7·0	27·8	∞	4·0	4	0·519
K^+C-4-S	3_2	7·0	27·8	∞	3·0	6	0·489
DNA	10_1	10·0	32·3	∞	3·0	10	0·413
DNA	10_1	10·0	32·3	∞	2·5	10	0·387
TMV	49_3	90·0	69·0	10·0	5·0	10	0·373
TMV	49_3	90·0	69·0	10·0	3·0	16	0·307
Pf1	27_5	30·0	75·6	10·0	5·0	6	0·458
Pf1	27_5	30·0	75·6	10·0	3·0	10	0·381

The K^+C-4-S data are based on a polycrystalline specimen (trigonal unit cell with $a = b = 13{\cdot}8$ Å and space group $P3_221$), and the other structures on non-crystalline specimens (continuous diffraction). References for these structures are given in the text.

m up to 20. The behavior of these quantities is illustrated in Fig. 1. Note that R_mS_m is almost constant except for very small *m*. For a particular fiber diffraction pattern, the N_m can be calculated, and the largest likely *R* factor calculated using (8) and the entries in Table 1.

3. Examples and discussion

The theory developed in the previous section is illustrated by calculating the largest likely *R* factors for four different structures. These examples represent the variety of types of structure that have been solved using fiber diffraction. Two of them, the potassium salt of the polysaccharide chondroitin-4-sulfate (K^+ C-4-S) (Millane, Mitra & Arnott, 1983) and a nucleic acid (Park, Arnott, Chandrasekaran, Millane & Campagnari, 1987), have rather small repeating units, and two, the helical virus TMV (Namba & Stubbs, 1985) and the bacteriophage Pf1 (Stark, Glucksman & Makowski, 1988), represent some of the largest structures solved by fiber diffraction. Three of the examples are based on continuous diffraction and one (chondroitin-4-sulfate) on diffraction from a polycrystalline specimen.

For non-crystalline structures, the number of overlapping terms at a particular cylindrical radius *R* in reciprocal space is determined by assuming that complex Fourier–Bessel structure factors G_n contribute to the diffracted intensity only for $n < 2\pi Ra + 2$ where $n \geq 2$ and *a* is the maximum radius of the molecule (Stubbs, 1989). The $n = 0$ and $n = 1$ terms contribute where $R \geq 0$ and $R > 0$, respectively. For polycrystalline specimens, the number of terms for each measurement is determined by the number of independent structure factors in the measured composite reflection. Reflections close to the meridian for which the molecular transform is very small (as determined by the above conditions on *n*) are excluded from the calculation. Meridional reflections are excluded since these are difficult to measure accurately and are not used in structure refinement.

The largest likely *R* factors for two maximum resolutions for each structure are listed in Table 2. These structures have a wide variety of diameters and symmetries and the largest likely *R* factors vary between 0·3 and 0·5 for typical values of the maximum resolution of the diffraction data. These represent typical largest likely *R* factors to be expected in fiber diffraction analyses, and show that values need to be calculated in individual cases. Inspection of Table 2 shows that *R* is strongly correlated with the maximum number of overlapping terms *M* on the diffraction pattern. Estimates using a weighted average for the R_m (Stubbs, 1989) give good approximations to largest likely *R* factors, although they are overestimated by 0·03 to 0·04 in these examples.

The theory developed here can be used to examine the dependence of *R* on parameters of a structure determination. The largest likely *R* factor was calculated for a hypothetical non-crystalline specimen with $a = 10$ and *c* repeat $c = 20$ Å as a function of diffraction data resolution $\rho_{\max}$ (for a structure with 10_1 helix symmetry), and helix symmetry u_1 (with a diffraction data resolution of 4 Å). The minimum resolution of the diffraction data was taken to be infinite. The results of these calculations are shown in Fig. 2. The largest likely *R* factor decreases with increasing resolution as the maximum number of overlapping terms increases. The *R* factor increases with increasing symmetry (increasing *u*) since this reduces the number of Fourier–Bessel terms at a particular position in reciprocal space. There are two features of the curve in Fig. 2(*b*) that are worthy of attention. For small *u*, *R* does not vary smoothly with *u* but appears to lie on two distinct curves, one for *u* even and one for *u* odd; and for large *u*, *R* is constant. The reasons for this behavior are as follows.

The number of values of *n* that satisfy the helix selection rule (Klug, Crick & Wyckoff, 1958) on layer lines $l = pu$ and $l = pu + u/2$, where *p* is an integer, is approximately half what it is on the other layer lines. Since there are no layer lines $l = pu + u/2$ when *u* is odd, the number of terms overall is larger, giving a smaller *R* factor when *u* is odd rather than even, as is evident in Fig. 2(*b*). This effect is more pronounced for smaller *u* as there are then more layer lines satisfying the above conditions. The effect is restricted to values of $u < u_0$ such that $l_{\max} = u_0/2$, where $l_{\max}$ is the largest layer-line number on the pattern, which gives

$$u_0 = 2c\rho_{\max} \qquad (11)$$

and R depends smoothly on u for $u > u_0$. In the above example $u_0 = 10$.

As the helix symmetry increases, the number of Bessel terms decreases until only one contributes on each layer line, so that $m = 1$ on the equator and $m = 2$ on the other layer lines. R therefore reaches a constant value,

$$R = R^{(\infty)}, \qquad u > u^{(\infty)} \qquad (12)$$

where $u^{(\infty)}$ can be estimated by determining when only one term contributes on the equator, which gives

$$u^{(\infty)} = 2\pi\rho_{\max}a + 2, \qquad (13)$$

$R_2 < R^{(\infty)} < R_1$ and $R^{(\infty)}$ depends on the number of layer lines. This situation corresponds to diffraction patterns dominated by single Bessel terms, which is unusual in high-resolution analyses of macromolecules, however. In the above example $u^{(\infty)} = 18$ and $R^{(\infty)} = 0{\cdot}627$. The proportion of the measurements contributed by the equator is approximately equal to $4/(\pi c\rho_{\max} - 4)$ so that

$$R^{(\infty)} \sim R_2 = 0{\cdot}586, \qquad c \to \infty \quad \text{or} \quad \rho_{\max} \to \infty. \qquad (14)$$

The fiber diffraction case for high helix symmetry and a small proportion of centrosymmetric reflections therefore approaches the single-crystal case. The above analysis is strictly valid only for integral helices, although the dependence of R on helix symmetry is qualitatively similar for non-integral helices.

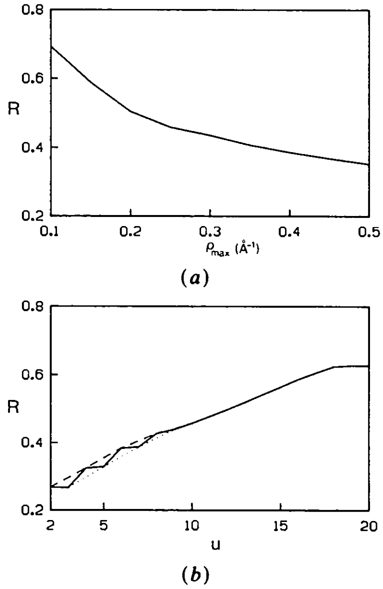

Fig. 2. Variation of the largest likely R factor R (for a structure with radius 10 Å and c repeat 20 Å) (a) with maximum resolution of the diffraction data $\rho_{\max}$ (for 10_1 helical symmetry) and (b) with helix symmetry u_1 (for a maximum resolution of 4 Å). In (b), the dashed and dotted curves are through points with u even and odd respectively.

4. Concluding remarks

The largest likely R factor for a general fiber diffraction pattern has been derived in terms of largest likely R factors for a single number of overlapping terms described previously. This allows accurate largest likely R factors for particular diffraction patterns to be calculated straightforwardly using the values listed in Table 1. Calculations for representative structures show that the largest likely R factor varies significantly with molecular diameter and symmetry and diffraction data resolution, and is typically between 0·3 and 0·5. These results also allow the effects of structural and diffraction parameters on R factors to be studied, and may be useful for further developments in fiber diffraction theory.

I am grateful to the US National Science Foundation for support (DMB-8606942) and Deb Zerth for word processing.

References

Klug, A., Crick, F. H. C. & Wyckoff, H. W. (1958). *Acta Cryst.* **11**, 199-213.

Millane, R. P. (1989). *Acta Cryst.* **A45**, 258-260.

Millane, R. P., Mitra, A. K. & Arnott, S. (1983). *J. Mol. Biol.* **169**, 903-920.

Namba, K. & Stubbs, G. (1985). *Acta Cryst.* **A41**, 252-262.

Park, H. S., Arnott, S., Chandrasekaran, R., Millane, R. P. & Campagnari, F. (1987). *J. Mol. Biol.* **197**, 513-523.

Stark, W., Glucksman, M. J. & Makowski, L. (1988). *J. Mol. Biol.* **199**, 171-182.

Stubbs, G. (1989). *Acta Cryst.* **A45**, 254-258.

Wilson, A. J. C. (1950). *Acta Cryst.* **3**, 397-399.