1. *Basic* PDB *Terms*

In a crystallographic (X-ray diffraction) experiment, the raw data consists of the positions and intensities of the reflections as measured in the diffraction pattern of the crystal. From these intensities, the structure-factor amplitudes can be calculated (roughly as the square root of the intensities). Once the phases of the structure factors are also known (i.e., once the "phase problem" has been solved), Fourier transformation of the structure factors provides a map, which is a three-dimensional matrix of numbers that represent the local electron density.[59] Where there are many electrons (and, hence, heavier atoms) the density is higher than in places where (on average) there are few electrons. It is now the task of the crystallographer to interpret the electron density in terms of a discrete atomic model.[60] This is typically an iterative process, in which the crystallographer (or in favorable cases even a computer program) builds a part of the model and then refines this. The refinement program will make small changes to the model by adjusting parameters such as the atomic coordinates, which improve the ability of the model to explain the experimental data. Simultaneously, geometric and other restraints and constraints are enforced onto the model to ensure that it is chemically reasonable. With an improved model, new maps can be calculated that may reveal further details, for example, previously missing or uninterpretable density for loops, ligand, solvent molecules, etc. The crystallographer can then add these. Simultaneously, the crystallographer should be on the lookout for possible errors in the current model and correct them if possible.[61]

Besides coordinates, atoms in the model typically have a "temperature factor" (also known as B factors or atomic displacement parameters) to model the effects of static and dynamic disorder in the crystal. Except at high resolution (typically, better than ~1.5 Å), where there are sufficient reflections to warrant refinement of anisotropic temperature factors (requiring six parameters per atom), temperature factors are usually constrained to be isotropic (requiring only one parameter per atom). The isotropic temperature factor of an atom is related to the atom's mean-square displacement. In most cases temperature factors provide a useful relative indication of the reliability of different parts of the model. If they are high, for example, for a lysine side chain, this usually means that little or no electron density was observed for the atoms in that side chain, and that the coordinates are therefore less reliable.

Figure 1 shows the atomic coordinate records of a crystallographically determined structure stored in the Protein Data Bank (PDB).[62] Figure 1a gives an example of crucial information in the REMARK records of PDB entries. Inspection of these notes and of a validation report (e.g., the WHAT IF report on the PDBREPORT web site or the PROCHECK report on the PDBsum web site) is highly recommended. In this case, the structure of crambin has been determined (PDB entry 1EJG). Crambin exists in two isoforms that differ in two residues (either Pro22/Leu25 or Ser22/Ile25), and both forms were present in the crystal. The two sequence heterogeneities have been modeled as alternative conformations for residues 22 and 25, but due to format restrictions, only one sequence is recorded in the sequence records.

Figure 1b shows a fragment of a PDB file from the same entry. The basic information about the atoms in the model is listed on "cards" (records, lines). These begin with ATOM for protein or nucleic acid components or HETATM for entities that are ligands, ions, metals, and solvent molecules. The second item on each line is simply a sequential index number of that atom. In the first line atom 136 is the amide nitrogen atom (N) of the valine (VAL) residue A8. "A" is the chain name, "8" the residue number. The "A" before the residue symbol "VAL" signifies that this atom is statically disordered. This means that this atom is observed in more than one location in the electron density, and the various instances are labeled "A", "B", "C", etc. Indeed, the third line in the figure contains the alternative location "B" of this atom. The three real numbers that follow the residue number—"6.382, 2.222, 13.070"—are the Cartesian coordinates (x, y, and z) of the atom in orthogonal Å. The fourth number is the occupancy of the position. This is a number between zero and one, which indicates the fraction of the amide nitrogen atom of valine A8 that occurs in this location. Here, the first conformation has been given an occupancy is 0.55, and line 3 shows that the alternative conformation B accounts for the remaining 0.45. Note that quite a few programs that read and process PDB files ignore alternative conformations completely. When the occupancy of ligands and solvent molecules is refined or set to a number less than one, this implies that they occupy the position in only a fraction of the molecules in the crystal, or for only a fraction of the time, or a combination of both. The fifth number, 1.92 in line 1, is the value of the isotropic temperature factor (B factor). Line 2 reveals that this atom has been modeled anisotropically, (this involves six parameters per atom which are listed on the ANISOU card), but the isotropic equivalent value is always listed as the fifth real number of the ATOM (or HETATM) card. At the end of each card the atomic symbol of the chemical element of the atom is listed, since this cannot always be deduced unambiguously from the atom's name.

An important parameter in crystallographic studies is the resolution of the data, which is expressed in Å, where lower numbers signify higher resolution. The higher the resolution, the more experimental data, and the more reliable (in terms of accuracy and precision) one may expect the resulting model to be. At high resolution (<1.5 Å) the model is probably more than 95% a consequence of the observed data.[63] However, at lower resolution (>2.5 Å), the modeling of details in protein structures is much more subjective than is widely appreciated.[64] This can be understood by calculating typical data-to-parameter ratios, that is, the ratio of the number of experimental observations and the number of adjustable parameters (atomic coordinates, parameters asso-

a)

```
HEADER    PLANT PROTEIN                                 02-MAR-00   1EJG
TITLE     CRAMBIN AT ULTRA-HIGH RESOLUTION: VALENCE ELECTRON DENSITY.
...
REMARK 999 PRO/SER22:LEU/ILE25 ISOFORMS ARE MODELLED
REMARK 999 AS ALTERNATE CONFORMERS IN COORDINATE RECORDS.
REMARK 999 BECAUSE OF FORMAT RESTRICTIONS, ONLY PRO22/LEU25
REMARK 999 ISOFORM IS REPRESENTED IN THE SEQUENCE RECORDS.
```

b)

```
ATOM    136  N  AVAL A   8       6.382   2.222  13.070  0.55  1.92           N
ANISOU  136  N  AVAL A   8      421    149    160     33    -23    -35       N
ATOM    137  N  BVAL A   8       6.695   2.072  13.037  0.45  1.88           N
ATOM    138  CA AVAL A   8       5.099   2.259  12.380  0.55  2.32           C
ANISOU  138  CA AVAL A   8      397    258    224     16    -22   -120       C
ATOM    139  CA BVAL A   8       5.471   2.048  12.164  0.45  1.94           C
ATOM    140  C   VAL A   8       5.208   3.386  11.373  1.00  2.63           C
ANISOU  140  C   VAL A   8      380    402    213     71    -42   -209       C
ATOM    141  O   VAL A   8       4.712   3.302  10.238  1.00  2.67           O
ANISOU  141  O   VAL A   8      378    434    200     39    -31   -178       O
ATOM    142  CB AVAL A   8       3.944   2.394  13.375  0.55  3.36           C
ANISOU  142  CB AVAL A   8      376    635    262    186      4   -244       C
ATOM    143  CB BVAL A   8       4.263   1.630  13.035  0.45  3.05           C
ATOM    144  CG1AVAL A   8       2.629   2.981  12.701  0.55  4.22           C
ANISOU  144  CG1AVAL A   8      326    856    420    159     34   -159       C
ATOM    145  CG1BVAL A   8       3.580   2.930  13.664  0.45  3.89           C
ATOM    146  CG2AVAL A   8       3.635   1.036  13.955  0.55  4.82           C
ANISOU  146  CG2AVAL A   8      771    684    376    230    -61   -385       C
ATOM    147  CG2BVAL A   8       3.136   1.077  12.028  0.45  4.17           C
ATOM    148  H  AVAL A   8       6.374   2.231  14.084  0.55  2.27           H
ATOM    149  H  BVAL A   8       6.648   2.059  14.045  0.45  5.91           H
...
END
```

Figure 1. a) An example of crucial information presented on REMARK records in PDB entries. b) Fragment of a PDB file from the same entry. The basic information about the atoms in the model is listed on "cards" (records, lines). For a complete description please refer to the text.

ciated with the temperature factors, and occupancies amongst others) in the model. For an average protein structure at a resolution of 2 Å, this ratio is slightly greater than two, but at ~2.7 Å it becomes less than unity. Whereas gross errors in the structure are unlikely to persist to the publication stage if the resolution is high, once the resolution becomes >2 Å, the balance shifts. Some published protein models appear to have been more determined by the crystallographer's imagination than by any experimental data.[63] In fact, in the 1980s the first reports of some of the "hottest" protein crystal structures, some of which were also prime drug targets, contained extremely serious errors.[65] Examples included HIV-1 protease, photoactive yellow protein, the small subunit of RuBisCO, D-Ala-D-Ala peptidase, ferredoxin, metallothionein, gene V binding protein, and the GTP-binding domain of Ha-ras p21.

Recently, the structure of a complex between botulinum neurotoxin type B protease and the inhibitor BABIM was published,[66] and the structure and experimental data were deposited in the PDB (entry 1FQH). However, subsequent critical analysis of the electron-density maps revealed that these did not support the placement of the inhibitor as stated in the earlier paper, and the structural conclusions based on it were withdrawn by the authors.[67]

Another trap to be aware of (and one that many crystallographers have fallen into) is the derivation of "high-resolution information" from low-resolution models. For instance, in a typical 3-Å structure the uncertainty in the position of the individual atoms can easily be 0.5 Å or more. Nevertheless, many such models have been described where hydrogen-bonding distances are listed with a precision (note: not accuracy!) of 0.01 Å (probably because the program that generated these distances used that particular precision) and solvent-accessible surface areas with a precision of 1 Å^2.

The ability of the model to explain the experimental data is usually assessed by means of the (conventional) R-value, which is defined in Equation (1).

$$R = \left(\sum \| F_{\mathrm{obs}} | - \mathrm{scale} \, | F_{\mathrm{calcd}} \| \right) / \left(\sum |F_{\mathrm{obs}}| \right) \qquad (1)$$

Here, F_{obsd} are the experimental structure-factor amplitudes, F_{calcd} are the structure-factor amplitudes calculated from the model, and the sums extend over all observed reflections. However, when more and more parameters are introduced into the model, the R-value can be made almost arbitrarily small (this is called "over-fitting the data"). In 1992 Brünger[68] introduced the concept of cross-validation in

crystallographic refinement, and with it the free *R*-value (*R*free), whose definition is identical to that of the conventional *R*-value, except that the free *R*-value is calculated for a small subset of reflections that is never used in the refinement of the model. The free *R*-value therefore measures how well the model predicts experimental observations that are not used to fit the model. Until a few years ago a conventional *R*-value below 0.25 was generally considered a sign that a model was essentially correct. While this is probably true at high resolution, it was subsequently shown for several intentionally mistraced models that these could be refined to deceptively low conventional *R*-values.[65,69] Brünger suggests a threshold value of 0.40 for the free *R*-value, that is, models with free *R*-values greater than 0.40 should be treated with caution.[70,71] Since the difference between the conventional and free *R*-value is partly a measure of the extent to which the model overfits the data (i.e., some aspects of the model improve the conventional but not the free *R*-value and are therefore likely to fit noise rather than signal in the data), this difference (*R*free $-$ *R*) should be small for the final model, ideally < 0.05.