Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

The Uppsala Electron-Density Server

Gerard J. Kleywegt,* Mark R. Harris, Jin-yu Zou, Thomas C. Taylor, Anders Wählby and T. Alwyn Jones

Department of Cell and Molecular Biology, Uppsala University, Biomedical Centre, Box 596, SE-751 24 Uppsala, Sweden

Correspondence e-mail: gerard@xray.bmc.uu.se

Received 23 January 2004
Accepted 2 June 2004

The Uppsala Electron Density Server (EDS; http://eds.bmc.uu.se/) is a web-based facility that provides access to electron-density maps and statistics concerning the fit of crystal structures and their maps. Maps are available for ~87% of the crystallographic Protein Data Bank (PDB) entries for which structure factors have been deposited and for which straightforward map calculations succeed in reproducing the published *R* value to within five percentage points. Here, an account is provided of the methods that are used to generate the information contained in the server. Some of the problems that are encountered in the map-generation process as well as some spin-offs of the project are also discussed.

1. Introduction

For experts and non-experts alike, macromolecular electron-density maps are the best representation of the crystallographic experiments that underpin the atomic models that are published and deposited. This is because models are just one crystallographer's subjective interpretation of the data and the maps (Brändén & Jones, 1990), reflecting that particular crystallographer's skills, experience and prejudices and possibly mistakes. Density maps, on the other hand, may reveal features that have not been interpreted as well as features for which an alternative interpretation may exist. Further, availability of an electron-density map enables users of a structure to assess the validity of claims made in the paper (active-site make-up, presence and conformation of bound ligands, nature of interactions *etc.*), and to carry out a proper assessment of the quality of the model (validation). For all these reasons, deposition of both model coordinates and experimental structure factors is mandatory according to the IUCr guidelines, which have been adopted by most journals that publish macromolecular crystal structures. However, even given the availability of model and data, only scientists with some knowledge of crystallography and with access to appropriate software are able to calculate electron-density maps. Therefore, we have undertaken to calculate such maps (in a uniform fashion) for all crystal structures in the Protein Data Bank (PDB) for which structure factors are available and to make the resulting maps (and statistics concerning the fit of model and data) available to the entire community of structure consumers through a server, the Uppsala Electron Density Server (EDS).

In this paper, we first review the history of the debate in the macromolecular crystallographic community concerning model and data deposition and briefly describe the current state of affairs. We then provide a detailed account of the methods that are used in the calculation of the EDS maps and

the problems encountered in this process. We further describe the kinds of files and information that are made available through the EDS server. Finally, a brief overview of the current state of the server is provided and possible future developments as well as some spin-offs of the project are discussed.

2. Deposition

Deposition of crystallographic data (coordinates and structure factors) has been hotly discussed since the late 1980s when the IUCr formulated a policy requiring deposition of such data. The first wave of discussions concerned the issue of whether or not deposition of coordinates should be made mandatory in conjunction with publication (Barinaga, 1989; Maddox, 1989; Koetzle, 1989). Eventually, most journals accepted the IUCr guidelines of the time (with *Nature* dragging its feet until 1996; Editorial, 1996). The IUCr guidelines allowed for a one-year delay on the release of coordinates and a four-year delay on the release of structure factors. In 1996, several groups of authors urged the crystallographic community to deposit structure factors for all their structures (Baker *et al.*, 1996; Jones *et al.*, 1996). Two years later, another round of discussions revolved around the issue of the allowed release delays (Wlodawer *et al.*, 1998; Editorial, 1998*a*, 1998*b*), with a number of journals eventually deciding to require immediate coordinate release upon publication (Campbell, 1998; Bloom, 1998). The IUCr, too, changed its guidelines after internal discussions (Baker & Saenger, 1999) and currently recommends deposition of coordinates and structure factors in the PDB, with release of coordinates upon publication, and of structure factors no more than six months after publication (Commission on Biological Macromolecules, 2000).

The mandatory deposition of structure factors is the next important issue that needs to be addressed by the community and the journals (but not necessarily the last issue: perhaps we should consider deposition of unmerged intensities or even raw diffraction images in the future). Fortunately, the community nowadays supports the notion of structure-factor deposition as judged by the record-high fraction in the year 2003 of structures deposited with the PDB for which structure factors were deposited as well (Kleywegt *et al.*, 2004). In 1995, only a third of all deposited crystal structures were accompanied by structure factors and in the period 1997–2002 this fraction hovered around two-thirds. However, in the year 2003 suddenly four out of every five crystal structure depositions included structure factors: a remarkable improvement and hopefully the beginning of a drive towards close to 100% structure-factor deposition. There are nevertheless considerable differences between different journals, with *Acta Crystallographica*, *EMBO Journal* and *Protein Science* reaching impressive structure-factor deposition levels of 90% or more, whereas *Nature*, *Science* and (disappointingly) *Biochemistry* are the only three journals for which fewer than two-thirds of the structures were accompanied by structure factors in the year 2003 (Kleywegt *et al.*, 2004).

The arguments against deposition (in particular of coordinates) and, later, in favour of extended release delays have been relatively few and most of them either do not apply any longer or can be addressed by postponing publication (at the risk of being scooped by competitors). With the sophistication of present-day refinement and model-building programs, as well as the speed of modern computers, the argument that time is needed to improve the accuracy of models has lost most of its validity. Delayed release of coordinates or data in order to file patent applications, to exploit the structure for ligand-design purposes or to reap more scientific rewards (follow-up studies *e.g.* of mutants and complexes) can be accomplished by delaying publication. The fear that others (bioinformaticians, theoreticians, competitors) might exploit a structure quicker than the scientist who determined it should encourage that scientist to broaden his expertise or to seek collaborations with experts in related areas. In some cases, low-resolution structures cannot be represented reliably by an all-atom model; in such (exceptional) cases, the IUCr guidelines provide for the deposition of a 'C^{α}-only' model (but accompanied by structure factors all the same). Unfortunately, this exception has been interpreted rather liberally at times. In a 1997 study of 'C^{α}-only' models (Kleywegt, 1997), fully 70% of all such models in the PDB at the time had been determined at better than 3.0 Å resolution (with 12% at 2.0 Å or better).

With respect to the deposition of structure factors specifically, some people have argued that they are superfluous since the refined temperature factors already provide an indication of the reliability of individual atoms. However, the arguments against this are many and sound. Firstly, temperature factors are not experimental data but model parameters that in addition are difficult to compare between different structures. Secondly, temperature factors are notorious for their role as 'error sinks'; they tend to account for much more than simply thermal vibration (*e.g.* unresolved disorder, partial occupancy, dynamic disorder, refinement artefacts such as inappropriate constraints or restraints, as well as possible errors in atom types, conformations *etc.*). This makes it essentially impossible to determine which factor(s) cause high temperature factors. Finally, temperature factors will never reveal any features in the density that have not been modelled or that could have been interpreted differently.

The arguments in favour of deposition (and immediate release) of coordinates and structure factors have been many. They can be clustered into a number of categories.

(i) Representation. The electron-density map is the best representation of the crystallographic experiment. It is rich in information and much less biased than an atomic model, revealing features that may not be included in the model or that are open to alternative interpretations. In order to calculate the density, however, both the model and the structure factors must be available.

(ii) Claim assessment. Without coordinates and structure factors, it is impossible to assess the validity of any claims made in the description of the structure, *e.g.* pertaining to the presence or conformation of bound ligands or even the correctness of the tracing. As Dickerson put it in 1989:

> 'I don't know of any other field of science where you are required to make public neither your data nor your results, only your commentaries'

(Barinaga, 1989).

(iii) Validation. This issue is related to the previous one but applies to all aspects of the model. Both coordinates and structure factors are needed for a thorough evaluation of the quality of the model and its usefulness for inclusion in high-quality databases or for further studies (*e.g.*, for modelling of complexes or mutants or as search models for molecular replacement). There are websites that run validation software on all coordinate entries in the PDB (Hooft *et al.*, 1996; Laskowski *et al.*, 1997) and report outliers. However, for many such outliers only combined use of coordinates and structure factors (or, rather, maps) enables one to classify them either as genuine (albeit unusual) features of the structure or as errors in the model (Jones *et al.*, 1996; Kleywegt, 2000).

(iv) Scientific progress. The availability of coordinates and structure factors facilitates and accelerates the progress of science. Follow-up studies by theoreticians, bioinformaticians, medicinal chemists, drug designers, biochemists, enzymologists, genetic engineers and other crystallographers are only possible if they have access to the structural results. There is also an element of reciprocity here: today a crystallographer may be a producer of a structure, but next month the same person may well be a consumer, in dire need of a search model for molecular-replacement purposes. Similarly, if the results are available upon publication, any interested scientist can immediately access and visualize them while reading the publication and the results can be presented to students in educational settings. Moreover, there is a very good case to be made for public funding agencies to require that results obtained with taxpayers' money be made available to the general public.

(v) Archival. If coordinates and structure factors are not deposited around the time of publication, the chances are that they will never be deposited and hence be lost forever to science. Phrased more selfishly, deposition is the ideal long-term archival solution for the results of crystallographic investigations. How many data sets have been lost to science because the postdoc who was responsible for the work left the institute or because the medium the data are stored on can no longer be read? Indeed, it has been noted that one of the most popular uses of deposited structure factors is for the crystallographers who did the experiment (or these persons' supervisors or successors) to be able to retrieve their own data which have been 'misplaced' in their own laboratories (Jiang *et al.*, 1999).

(vi) Databases. In principle, making coordinates public could be left to individuals or journals, *e.g.* through websites. However, in the macromolecular-structure world we are fortunate to have one single uniform database in the form of the Worldwide PDB (Berman *et al.*, 2003). Whereas a motley collection of files scattered over many physical locations is better than nothing, its usefulness pales compared with that of a central database. Database-wide ('mining') studies allow for comparisons (of structures, electron density, temperature factors, water structure *etc.*), classifications (*e.g.* of folds or complexes), derivation of statistics (*e.g.* regarding geometrical and other structural features to be used for model building, refinement or validation) *etc*. Moreover, as new methodology is developed a database guarantees that a wide selection of cases is available for testing purposes.

(vii) Vanity. Although this argument to our knowledge has never been expressed publicly, surely it must appeal to any crystallographer's sense of professional pride if their structures are actually used by colleagues, coworkers, competitors, teachers, students, popular science magazines *etc.* In addition, structures and data that can be readily accessed are likely to receive more citations than their unavailable brethren and will eventually survive longer in the annals of science.

3. Generating the maps

The process of calculating the electron-density maps for EDS involves downloading the coordinate and structure-factor files from the PDB (Berman *et al.*, 2000), conversion of the CIF-format reflection files to *CCP*4 (MTZ) format, modification of the coordinate files to make them suitable for processing with *REFMAC* (Murshudov *et al.*, 1997), calculation of structure factors and map coefficients with *REFMAC*, calculation of σ_A-weighted (Read, 1990) and $F_{\rm calc}$ maps with *CCP*4 (Collaborative Computational Project, Number 4, 1994) programs, calculation of real-space R values and other residue-based statistics with *MAPMAN* (Kleywegt & Jones, 1996*a*) and the generation of files that can be downloaded by EDS users. Every now and then, this process is carried out from scratch for all entries. For this we use a Linux-based computer cluster with seven nodes, which allows the calculations to be performed in $\sim$3 d. In addition, the server is updated automatically every weekend, when new and updated coordinate and structure-factor files are downloaded from the PDB and processed. The update process is carried out by a C-shell script, whereas the map calculations for individual entries are performed by a Perl script that carries out the following steps.

(i) The CIF-format reflection file retrieved from the PDB is converted into an ASCII file containing h, k, l, $F_{\rm obs}$ and $\sigma(F_{\rm obs})$ using a 'jiffy' program, *CIF2TEXT*. This program also creates a small text file (*xxxx*.sfdat, where '*xxxx*' is the PDB code of the entry) with the number of reflections, the calculated resolution *etc*. If a CIF file contains intensities instead of amplitudes, they are converted to amplitudes using $F = I^{1/2}$ and $\sigma F = \frac{1}{2} I^{-1/2} \sigma(I)$. However, if a reflection file stems from a neutron, electron or fibre diffraction or electron-microscopy experiment, the entry is not processed any further.

(ii) The resolution and unit-cell constants of the entry are retrieved from the PDB-supplied index file resolu.idx. If the resolution quoted in the PDB file differs substantially (more than 0.75 Å) from that calculated by *CIF2TEXT*, the processing of this entry is aborted. If the difference is large, but not outrageous, a warning message is generated.

(iii) *DATAMAN* (Kleywegt & Jones, 1996*a*) is run to limit $\sigma(F_{\rm obs})$ to reasonable (positive) values, to remove reflections

with non-positive F_{obs}, to carry out some sanity checks and to write the resulting reflection data to a new ASCII file. If one or more of the sanity checks fail, the entry is not further processed. Current sanity checks detect cases where (almost) all Miller indices h or k or l are identical and cases where (almost) all values of F_{obs} or $\sigma(F_{obs})$ are equal; failure to pass any of these tests tends to signal a corrupted reflection file. Another sanity check detects cases where the average ratio $F_{obs}/\sigma(F_{obs})$ is less than one, suggesting that the labels of F_{obs} and $\sigma(F_{obs})$ may have been swapped in the CIF file.

(iv) The overall atomic radius that is to be used in the real-space fit calculations (carried out by *MAPMAN*; see below) is derived using the following rules: if the resolution is less than 0.6 Å, the radius is set to 0.7 Å. If the resolution is greater than 3.0 Å, the radius is set to half the resolution (in Å units). For resolution values in between, the radius is set to: 0.7 + (resolution − 0.6)/3.0 Å. The map border around the molecule (used by *MAPMASK*) is set to this radius value plus 3.0 Å.

(v) The coordinate file is analysed to see if strict implicit NCS is present. If this is the case, the operators are retrieved (from the MTRIXn records) to enable expansion of the contents of the coordinate file under full NCS.

(vi) Incomplete 'C^{α}-only' (or, rather, 'C^{α}-mostly') entries are identified by counting the number of C^{α} atoms. If there are more than 30 C^{α} atoms and the C^{α} atoms make up at least one-third of all the atoms, then the entry is not further processed.

(vii) The space group is retrieved, and some space group names are modified to conform to *CCP*4 naming conventions, *e.g.* *P*1211 becomes *P*21, *H*3 becomes *R*3 *etc*.

(viii) The ASCII reflection file obtained from *DATAMAN* is converted to MTZ format using the *CCP*4 program *F2MTZ*.

(ix) The *CCP*4 *uniqueify* script is run to produce a new MTZ file and to calculate the completeness of the data.

(x) A 'jiffy' program (*FILPDB*) is used to modify the coordinate file of the structure so that it can be processed by *REFMAC*. This program carries out the following operations: all HETATM records are renamed to ATOM records; all REMARK 290 records are removed; if charges are present in columns 79–80 they are removed; deuterium atom names in columns 77–78 are replaced by hydrogen; any non-alphabetic characters in columns 77–78 are replaced by spaces; any UNK/UNX atoms and residues are removed (because it is not clear *a priori* what kind of scattering factors would need to be used); any 'AD1', 'AD2', 'AE1', and 'AE2' atom names in Asp, Asn, Asx, Glu, Gln and Glx residues are replaced by their normal names (assuming Asn for Asx and Gln for Glx); rhombohedral *H* space groups are renamed to *R*; any scales defined on SCALEn records are applied to all coordinates; element names are added if they are not given (only for HETATM atoms; using the information on the FORMUL records); if implicit NCS is present, the appropriate operators are applied to generate the complete asymmetric unit.

(xi) *REFMAC* is run to apply anisotropic scaling, to perform bulk-solvent correction and to calculate structure factors and map coefficients (for the σ_A-weighted maps).

(xii) σ_A-weighted $2mF_{obs} - DF_{calc}$ and $mF_{obs} - DF_{calc}$ maps as well as an F_{calc} map are generated with the *CCP*4 program *FFT*.

(xiii) *MAPMASK* is run to cut out the maps around the molecule(s) present in the coordinate file.

(xiv) *MAPMAN* is used to convert the $2mF_{obs} - DF_{calc}$ and $mF_{obs} - DF_{calc}$ maps into *O*-style DSN6 maps (using one byte per grid point to store the electron-density values). Before scaling the floating-point electron-density values to single-byte integers, the dynamic range of the density values is first capped to curb spikes in the density (at +5 e Å^{-3} for both types of map, as well as at −2 e Å^{-3} for $2mF_{obs} - DF_{calc}$ maps and at −5 e Å^{-3} for $mF_{obs} - DF_{calc}$ maps). Subsequently, a linear transformation is used to map the density values to single-byte integers in the range 0–255.

(xv) Residue-based statistics (real-space *R etc.*) are calculated using *MAPMAN*. The log files produced by *MAPMAN* in this step, together with the *O*-style maps, constitute the core data in EDS.

(xvi) The *MAPMAN* log file serves as input to a 'jiffy' program (*STAT2O*) that generates a file with *O* data blocks and *O* macros.

(xvii) At each step, statistics are being added to the *xxxx*.sfdat file. In addition, notes, warnings and errors are collected in a file called ERRORLOG.

(xviii) The entry's directory is cleaned up by removing all intermediate and scratch files. At this stage, the top-level C-shell script takes over again. It creates a date-stamp file containing the current time and date and an *O* macro to read and draw the molecule and the map *etc.*

4. Problems

Despite our best efforts, there are still a large number of PDB entries for which coordinates and structure factors are available, but for which we are unable to calculate structure factors such that the reported *R* value is reproduced (to within five percentage points). These failures may be caused by problems with the coordinate files, problems with the structure-factor files, or limitations of the software that we use. Below, we describe the three categories of problems that we encounter and some of the most common causes of these problems.

(i) Uninterpretable structure-factor files. These are usually files that have not yet been converted into CIF format by the RCSB (and in some cases files from fibre diffraction or electron-microscopy experiments). We have been able to convert a number of problematic files ourselves, but there remain files for which it is impossible to guess which column represents F_{obs} (if any).

(ii) Failed map calculations. This signals that somewhere along the line the map-calculation process broke down. This can be owing to the following:

(1) Entries that pertain to neutron, electron or fibre diffraction or electron-microscopy experiments.

(2) Problems with the CIF file. The sanity checks carried out by *DATAMAN* catch a number of these problems. Examples are data sets in which all h or k or l Miller indices are equal

(usually zero), data sets for which the average value of the ratio $F_{obs}/\sigma(F_{obs})$ is less than one (suggesting that their labels were swapped in the CIF file) and data sets that contain fewer than 250 reflections (some CIF files do not contain a single reflection).

(3) A large difference between quoted and calculated resolution. This may be owing to the wrong structure-factor file having been deposited, 'silly' reflections having been included (*e.g.* H = 999 to signal an end-of-file), processing problems at the PDB that have corrupted the structure-factor file, an inconsistency between the indexing of the reflections and the unit-cell parameters (*e.g.* in $P2_12_12_1$ when a long and a short axis have been swapped) or a conscious decision by the crystallographer to use a much lower resolution cutoff for refinement than was used for data processing but nevertheless to deposit the entire data set. In one puzzling case (PDB code 1jzp), the discrepancy is caused by the fact that the coordinate entry belonging to the structure factors is actually an NMR model.

(4) Problems with the CIF file labels (*e.g.* typing errors or the use of labels for F_{obs} or I_{obs} that are not recognized by *CIF2TEXT*).

(5) Space-group problems. Unusual space groups or settings (*e.g.* *A*1, *A*2, *B*2, *P*112, *I*2) usually cause the *uniqueify* script to fail. Space group 'P 21 21 2 A', on the other hand, will cause *MAPMASK* to fail.

(6) Entries that consist largely or exclusively of C^{α} atoms.

(7) Big structures. For example, viruses can be so big that the *MAPMASK* program fails. Also, our current version of *REFMAC* cannot handle more than 100 000 atoms, thus rendering map calculations unfeasible for ∼20 PDB entries. Even when the map calculations succeed, however, virus structures are a particular concern. It would be best if NCS-averaged maps could be made available for these entries, especially for those with relatively low structure-factor completeness. We hope to resolve this problem in discussion with the depositors.

(8) There are a few bugs in *REFMAC* that affect a small number of entries.

(iii) A large difference between quoted and calculated R values. This is the most difficult class of entries to analyse: the map calculations succeed technically, but the R value reported by *REFMAC* differs by more than five percentage points from that reported in the PDB file. Some of the possible causes that we have identified in the past are as follows.

(1) Different methodology for the structure-factor calculations. For instance, EDS uses *REFMAC* to perform a bulk-solvent and anisotropic correction (the authors may not have used such corrections or they may have carried them out with different software). Further, EDS does not apply any special corrections such as for twinning (not yet supported by *REFMAC*), whereas the authors may have applied such corrections. Also EDS does not handle unmerged reflection data properly.

(2) Different sets of reflections. EDS uses all data found in the structure-factor file, whereas the authors may have used one or two resolution cutoffs, an $F_{obs}/\sigma(F_{obs})$ cutoff *etc.* Further EDS uses all reflections, whereas the authors may have kept work and test reflections separated. On the other hand, EDS uses amplitudes instead of intensities, so if intensities were used and deposited, reflections with non-positive intensities will be discarded by EDS.

(3) Problems with the PDB coordinate file. For instance, the R value reported in the PDB file may be wrong or not quoted at all or reported in a non-standard fashion, the coordinates may not be those described in the paper or deposition form (*e.g.* waters were not deposited), the reported unit-cell parameters may be incorrect (in one case, an error of 1.0 Å had increased the R value from 0.27 to 0.39) or the definition of any NCS relationships may be non-standard or incorrect (*e.g.* implicit NCS may have been labelled as being explicit or *vice versa*).

(4) Problems with the structure-factor file. For example, calculated instead of observed structure-factor amplitudes may have been deposited, the CIF file is said to contain amplitudes whereas it does in fact contain intensities (or the other way around), whatever has been labelled as F_{obs} is in fact something completely different, the structure-factor file and the coordinate file do not match [*e.g.* structure factors belong to a (different) complex or mutant] or the deposition script of *CNS* was used which modifies F_{obs} to correct for anisotropy, which is undesirable and leads to rejection of reflections with negative 'F_{obs}' by our software.

(5) Problems with or limitations of the software that is used (and that we are possibly not even aware of). For example, there exist depositions that are so large that they have been spread out over multiple PDB entries.

5. The EDS web-server

The electron-density server can be accessed through the URL http://eds.bmc.uu.se/.

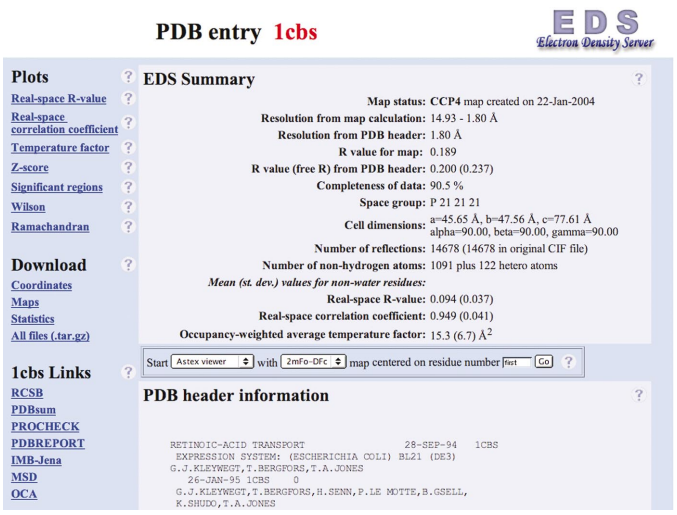

Figure 1
Example of an entry page in EDS (1cbs; Kleywegt *et al.*, 1994) showing the available information, plots, files and external links discussed in the text.

An entry can be accessed directly by providing its PDB code. Alternatively, a rudimentary keyword search can be carried out. In addition, some database centres provide links to EDS, *e.g.* the RCSB PDB site (Berman *et al.*, 2000) and the IMB Jena Image Library of Biological Macromolecules (Reichert *et al.*, 2000) and the search facilities available at these centres can therefore also be used to locate a certain EDS entry. Information on how to link to specific EDS entries is provided on the EDS website. For each EDS entry we provide the following information (Fig. 1).

(i) General information and overall statistics. A number of crystallographic statistics are listed (resolution, R values *etc.*), some of which compare the values reported in the parent PDB file and those obtained in the EDS map-calculation process. Further, some information from the header of the parent PDB file is listed (*e.g.* the system that was studied, the authors and any literature references).

(ii) Interactive display of model and map. A form is presented that can be used to launch a Java-based viewer program to inspect the model and the density. The user can select which viewer program to use, which map to display and on which residue the viewer should initially centre. At present, we support two viewers, one being a simple viewer developed in-house (MRH, unpublished results) and the other being the EBI-version of the *AstexViewer* (Hartshorn, 2002), the latter being the default viewer. The *AstexViewer* (Fig. 2*a*) has more controls, can display multiple maps, is faster and more user-friendly and is the same viewer as is used by the EBI tools for the MSD databases (Boutselakis *et al.*, 2003). On the other hand, the Uppsala viewer (Fig. 2*b*) has slightly better rendering, is simpler and its controls are more similar to those in *O*.

(iii) Plots. Plots of various statistics can be generated on-the-fly. Clicking on a residue in some of these plots will start up the Java-based viewer centred on that residue, enabling inspection of the residue, its environment in the structure and the local electron density. The following plots can be inspected for every entry:

(1) Real-space R value (RSR): one plot per chain (Fig. 3) as a function of residue number (only for protein and nucleic acid chains). The values for all water molecules are sorted and plotted by increasing RSR values. The values for hetero entities are listed in a table at the bottom of the page. The real-space R value was introduced by Jones *et al.* (1991). Two maps, one 'observed' (in this case, a σ_A-weighted $2mF_{\rm obs} - DF_{\rm calc}$ map) and one 'calculated' (in this case, an $F_{\rm calc}$ map; note that

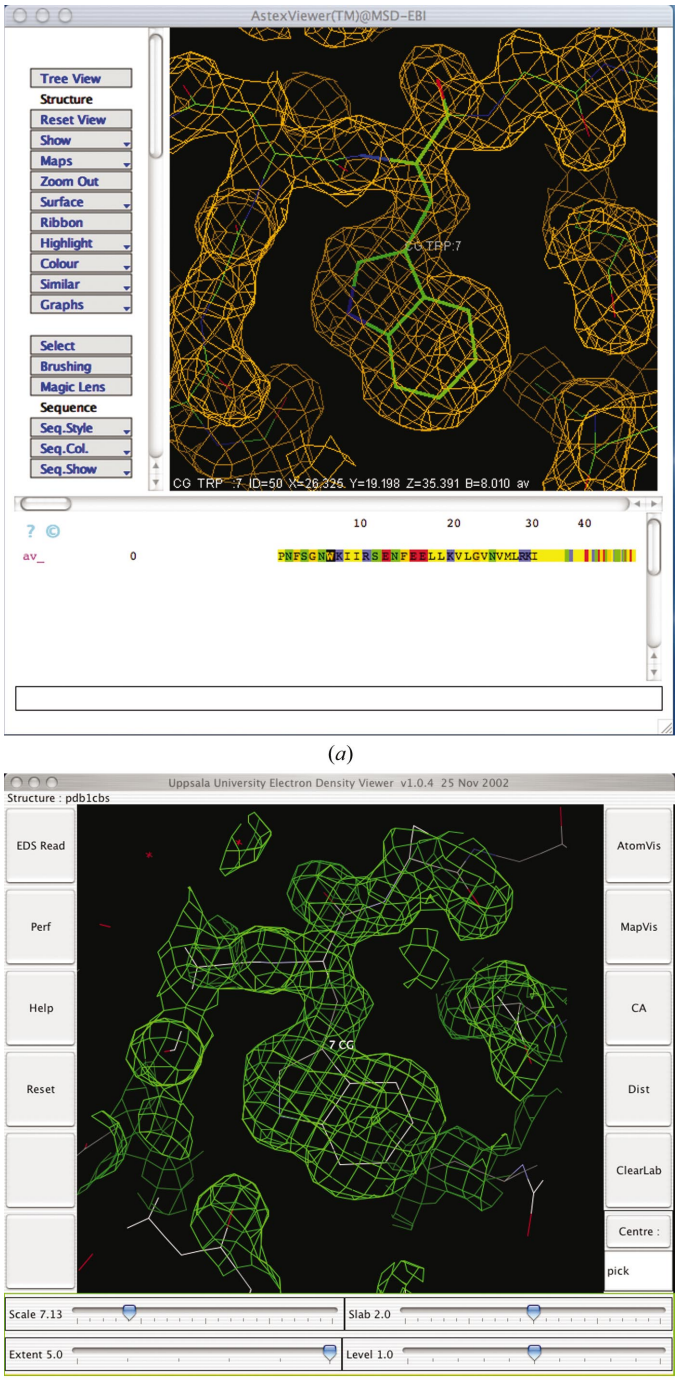

Figure 2
(*a*) Example display of the *AstexViewer* showing model and EDS density for residue Trp 7 in entry 1cbs. (*b*) The same data as in (*a*), but visualised here with the Uppsala viewer.

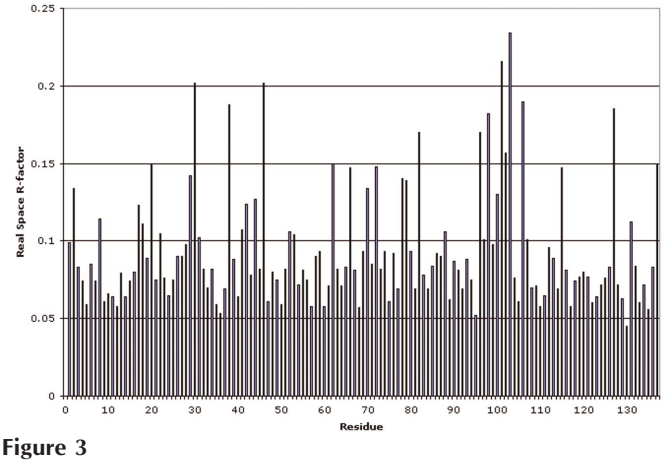

Figure 3
Example of a residue-based plot of real-space R values (for entry 1cbs). The bar of every residue is clickable in the browser and will launch the density viewer, load and display the appropriate model and map and centre on the selected residue.

this is different from the implementation of Jones *et al.*, 1991), are needed. For every residue, all density values within a certain radius of each of its constituent atoms (taking any and all alternative conformations into account) are compared in both maps. The RSR-value is then defined as: RSR = $\sum |\rho_{obs} - \rho| / \sum |\rho_{obs} + \rho_{calc}|$, where the sums extend over all the density values considered. The radius that is used depends on the resolution of the data (see above).

(2) Real-space correlation coefficient (RSCC). The calculation of RSCC values is similar to that of the RSR values, except that the linear correlation coefficient between the two density arrays is calculated for every residue. This value will always lie between −1.0 (perfect anti-correlation) and +1.0 (perfect correlation), with values close to zero signifying lack of any correlation. Use of the correlation coefficient has the advantage that the observed and calculated density need not be scaled together. A minor drawback is that weak density that has the proper intensity distribution will get a high score (sometimes this is noticeable for water molecules with weak but spherical density).

(3) Temperature factor. The occupancy-weighted average temperature factor for every residue is plotted or listed. This is defined as $\langle B_{iso} \rangle = (\sum B_{iso} \times Q)/(\sum Q)$, where the sums extend over all atoms in the residue (including any and all alternative conformations) and Q represents occupancy.

(4) Z scores. The RSR value of each residue (proteins and nucleic acids only) is used to calculate a resolution-dependent Z score: $Z = (\mathrm{RSR} - \langle \mathrm{RSR}_{resolution} \rangle)/\sigma(\mathrm{RSR}_{resolution})$, where $\langle \mathrm{RSR}_{resolution} \rangle$ is the average, and $\sigma(\mathrm{RSR}_{resolution})$ is the standard deviation of the RSR values of all residues of the same type (*e.g.* arginine) in the resolution range that the structure lies in (*e.g.* 1.4–1.6 Å). As an example, Table 1 shows database-wide RSR statistics pertaining to leucine residues. A large, positive spike in a Z-score plot implies that a residue has an RSR value that is considerably worse than that of the average residue of the same type in structures determined at similar resolution.

(5) Significant regions. This plot only shows residues for which the Z score is greater than 2.0.

(6) A Wilson plot, showing the relation between the logarithm of the intensity of the reflections and the resolution.

(7) A Ramachandran plot, generated using the definition of core regions of Kleywegt & Jones (1996*c*) as implemented in the program *MOLEMAN*2. Plots are produced for every protein or peptide chain separately. Below the plot a list of outlier residues is shown. Clicking on the name of an outlier residue will start the Java viewer and centre on that residue.

(iv) Downloadable files. The following items can be downloaded for every entry.

(1) Coordinates can be downloaded in PDB format.

(2) The electron-density maps ($2mF_{obs} - DF_{calc}$ and $mF_{obs} - DF_{calc}$). The maps are stored in the compact DSN6 format, which can be read by the crystallographic modelling and graphics program *O* (Jones *et al.*, 1991) and also by *Swiss PDB Viewer* (Guex & Peitsch, 1997), *WHAT IF* (Vriend, 1990) and other molecular graphics programs. However, the maps can be downloaded not only in DSN6, but also in *CCP*4, *CNS* (Brünger *et al.*, 1998) and *EZD* (GJK & TAJ, unpublished results) format (note, however, that the precision of these maps is still 1 part in 256 as dictated by the DSN6 format).

Table 1
Example of residue- and resolution-specific real-space *R*-value statistics (for leucine residues; calculated in December, 2003).

Resolution (Å)	⟨RSR⟩	σ(RSR)	Observations
15.0–3.0	0.2226	0.0977	31743
3.0–2.8	0.1915	0.0768	35200
2.8–2.6	0.1779	0.0694	40144
2.6–2.4	0.1602	0.0614	56775
2.4–2.2	0.1429	0.0574	59352
2.2–2.0	0.1270	0.0506	73999
2.0–1.8	0.1148	0.0468	73236
1.8–1.6	0.0983	0.0393	42847
1.6–1.4	0.0864	0.0330	18131
1.4–1.2	0.0822	0.0310	5553
1.2–0.0	0.0681	0.0254	3446

(3) Statistics. This is the listing of the real-space *R* values, occupancy-weighted average temperature factors, *etc*. for all residues, plus some overall statistics (the log file produced by *MAPMAN*).

(4) All files. This enables the user to download a compressed archive file that contains the model and maps, the reflection file from *REFMAC*, *O* macros, statistics *etc.* Some of these files are only useful if one has access to the program *O*, whereas others provide more information about the results of the calculations. They may be of use to expert users who want to carry out further crystallographic computations.

(v) External links. For every entry, there are links to other web-based services that provide information about that particular entry. At present, EDS provides links to RCSB (Berman *et al.*, 2000), the home of the PDB, providing access to the original PDB entry as well as lots of related information and links to further databases; PDBsum (Laskowski *et al.*, 1997), providing a one-page summary of the entry, plus many useful figures and links; *PROCHECK* (Laskowski *et al.*, 1993) analysis (part of PDBsum); *PDBREPORT* (Hooft *et al.*, 1996), providing a *WHAT IF* validation report for the entry; IMB Jena Image Library of Biological Macromolecules (Reichert *et al.*, 2000), for images of the entry; MSD (Boutselakis *et al.*, 2003), EMBL–EBI's macromolecular structure database; and OCA (http://bip.weizmann.ac.il/oca-docs/oca-home.html), a PDB front-end.

6. Preliminary results, spin-offs and future developments

As of the end of November, 2003, EDS comprised 23 267 PDB entries, of which 19 864 were crystal structures. For 10 751 of these (54%), structure factors were available. For 104 (1%) of these entries we were unable to calculate maps, whereas for 9394 (87%) of them the *R* value calculated by us agrees within five percentage points with the one reported in the PDB entry. These numbers imply first and foremost that for almost half of all deposited crystal structures no experimental data has been

deposited. Unless a major effort is made now by the responsible scientists, we have to fear that this data will be lost to science forever. To help crystallographers identify for which of their entries (if any) structure factors have yet to be deposited, we provide an easy-to-use web-based form (http://eds.bmc.uu.se/eds/eds_sos.html). The second conclusion is that a for a sizeable number of entries (more than 1200 at present) straightforward calculations using the deposited coordinates and structure factors are not sufficient to reproduce the published R values to within a reasonable margin (five percentage points). Although there are certainly cases where our software is simply not sufficiently advanced, in many of the cases where we or the original depositors have been able to pinpoint the source of the problem it has involved errors (often of a book-keeping nature) that were introduced at the deposition stage. To track down the source of the problems in the remaining cases, help from the depositors is invaluable. Authors who find that their entries are not represented in EDS may, as a first step, want to download their own files from the PDB and attempt to reproduce their published R values. If these attempts fail, it should not be too difficult for them to track down the problem by looking for discrepancies between the files that were actually used during structure refinement and those that were downloaded from the PDB. Many of the problems are trivial and easily correctible by the authors (but usually not by anyone else!) and may be due to typographical errors, swapping of indices or cell constants, or mixing up of related files. That these problems have not been detected previously is because the EDS project is probably the first in which a systematic effort is made to calculate electron-density maps for all of the more than 10 000 crystal structures for which structure factors are available. As a community we need to make an effort to fix problems in the existing database entries and we need to do it sooner rather than later, while the original files still exist (on media that can still be read with modern equipment) and while the people who did the work are still around. For the future, problems can be prevented only by making map calculations an integral part of the data-deposition process. To this end, we have been working with the MSD group to make EDS-style calculations part of the deposition process at their site. Nowadays, when a crystallographer deposits a model and structure factors at the EBI site, the EDS calculations are carried out automatically, summary statistics presented and the resulting files are made available to the depositor.

Thanks to the fact that EDS now contains more than 9000 electron-density maps, all calculated in a consistent fashion, we have a large set of statistics pertaining to maps waiting to be 'mined'. For example, we investigated what factors the unit-cell r.m.s. density level ('σ-level') depends on. Our initial assumption was that it would be correlated with the solvent content of the parent crystal, but when this statistic was calculated for a set of entries, the correlation turned out to be poor. We should not have been surprised to find that the strongest correlation was in fact with the occupancy-weighted average temperature factor (averaged over all non-water entities so as to yield a single value for every entry). The inverse correlation (Fig. 4) suggests that it could be advantageous to use (dynamically and automatically) variable contouring levels during model inspection and rebuilding, where the locally averaged occupancy-weighted temperature factor determines the appropriate contouring level.

In addition to the overall statistics, EDS also provides detailed statistics concerning the real-space fit of all residues in more than 9000 crystal structures allowing comparative and retrospective studies, *e.g.* concerning the fit of ligands, the reliability of water molecules *etc.* To date, however, we have only used these statistics to identify some cases of poorly fitting molecules to use as educational examples in lectures.

The residue-specific real-space fit statistics (such as those in Table 1) are an additional valuable by-product of the EDS project. They will make it possible to use residue- and resolution-specific RSR-cutoff values in validation procedures (*e.g. OOPS*2; Kleywegt & Jones, 1996*b*). More importantly, they can be employed in automatic rebuilding programs and protocols such as *ARP*/*wARP* (Perrakis *et al.*, 1999), for instance by using heuristic rules such as 'remove or rebuild all residues for which Z(RSR)> 2'. A prototype program, *ELAL* ('ELectronic ALwyn'), that applies such heuristics has been written (GJK, unpublished results) and will be integrated into a future version of *ARP*/*wARP* (Cohen *et al.*, 2004).

The availability of a large number of maps also makes it possible to study the phenomenon of register errors. This type of model-building error occurs when one or more residues are skipped or inserted into the model to render the sequence and the model out of sync (Kleywegt *et al.*, 1996; Jones & Kjeldgaard, 1997). At present it is impossible to determine how common such errors are in deposited models since at least two models of the same molecule are needed to detect any register shifts and the density is needed for both to determine if either of them is in error. More importantly, there are no (combinations of) diagnostics that are known to be specifically suited to detecting such errors in models prior to their deposition (especially in the absence of comparison models). We have

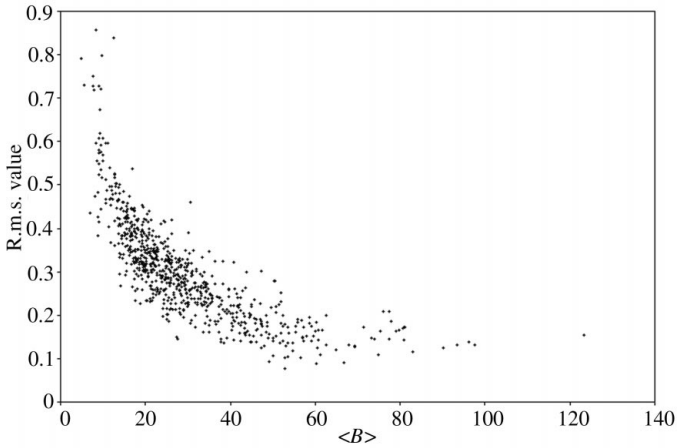

Figure 4
Plot of the unit-cell r.m.s. density level ('σ') of 1000 σ_A-weighted $2mF_{obs} - DF_{calc}$ maps (units of e Å^{-3}) as a function of the occupancy-weighted average temperature factor (averaged over all non-water entities; units of Å^2). The linear correlation coefficient is -0.75.

therefore undertaken a study of sets of crystal structures of the same molecule between which register shifts (not necessarily errors) occur and will try to correlate these to coordinate-based and map-based validation statistics (H. Hansson & GJK, unpublished results].

Finally, the work on EDS has probably made a small contribution to improving the quantity, quality and integrity of the structural database as a whole. We have identified (and sometimes resolved) a number of problems with structure-factor files that had not come to light previously. Further, quite a few crystallographers have used the web form that we provide to identify which of their models did not have structure factors deposited (in some cases leading to the deposition of several dozen old data sets by a single crystallographer). Others have examined their deposited models and structure-factor files and identified and corrected mistakes that were made during the original deposition process.

As for future developments, we are constantly working on improving the methods that are used to calculate the maps, and on trying to identify causes of failed map calculations. With the help of the crystallographic community we hope to improve the 87% success rate. At a later stage, we will also attempt to incorporate maps that are provided by the crystallographers themselves (experimental maps, NCS-averaged maps *etc.*). We are also working on a simple server with which crystallographers will be able to execute the EDS-style calculations prior to deposition. Finally, in the long term the community will probably have to face the issue of whether the structural database should be static or dynamic. As methodology improves, it seems likely that re-refinement of older models (either on a case-by-case basis, or as one large-scale project) might provide better models and, hopefully, increase our understanding of the chemistry and biology of the molecules under study.

7. Conclusions

At present, structure factors are available for only 54% of all crystal structures deposited in the PDB. Unless the community makes a serious effort now, we must assume that the remaining data is lost to science forever. Fortunately, structure-factor deposition has become vastly more common in recent years, with 78% of all crystal structures deposited in the year 2003 being accompanied by the corresponding data set.

Electron-density maps are the best representation of the crystallographic experiment (both for experts and non-experts). However, their computation requires crystallographic expertise and access to the proper software. Therefore, we have calculated maps for more than 9000 macromolecular crystal structures in the PDB and make these available through an internet-based server, the Uppsala Electron Density Server (EDS). In doing so, a 'dead' collection of structure-factor files has been transformed into a publicly accessible collection of thousands of maps that can be inspected, scrutinized and admired by experts and non-experts alike.

We thank Johan Hattne (Uppsala University) for his help in implementing the interactive map viewing in EDS, Mike Hartshorn (Astex Technology) and Tom Oldfield (EBI) for developing and modifying the *AstexViewer*, Jawahar Swaminathan (EBI) for implementing the EDS software at EBI and for many corrections to legacy structure-factor files in the PDB archive, and Kim Henrick (EBI) for useful discussions and the EBI–EDS collaboration. The *AstexViewer* software is used in EDS by permission and includes code developed by Astex Technology Limited, UK. In its initial stages, EDS was supported in part by EU contract NO. CT96-0189. The Swedish Natural Science Research Council (NFR, now VR) and the Wallenberg Foundation (through the Linnaeus Centre for Bioinformatics in Uppsala) have supported later developments. Current support is provided by EU contract No. QLRT-2001-00015. GJK is a Royal Swedish Academy of Sciences (KVA) Research Fellow, supported through a grant from the Knut and Alice Wallenberg Foundation. He is supported by KVA, the Swedish Structural Biology Network (SBNet), Uppsala University and its Linnaeus Centre for Bioinformatics.

References

Baker, E. N., Blundell, T. L., Vijayan, M., Dodson, E., Dodson, G., Gilliland, G. L. & Sussman, J. L. (1996). *Nature (London)*, **379**, 202–202.
Baker, E. N. & Saenger, W. (1999). *Acta Cryst.* D**55**, 2–3.
Barinaga, M. (1989). *Science*, **245**, 1179–1181.
Berman, H., Henrick, K. & Nakamura, H. (2003). *Nature Struct. Biol.* **10**, 980.
Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.
Bloom, F. E. (1998). *Science*, **281**, 175.
Boutselakis, H. *et al.* (2003). *Nucleic Acids Res.* **31**, 458–462.
Brändén, C.-I. & Jones, T. A. (1990). *Nature (London)*, **343**, 687–689.
Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.
Campbell, P. (1998). *Nature (London)*, **394**, 105.
Cohen, S. X., Morris, R. J., Fernandez, F. J., Ben Jelloul, M., Kakaris, M., Parthasarathy, V., Lamzin, V. S., Kleywegt, G. J. & Perrakis, A. (2004). *Acta Cryst.* D**60**, 2222–2229.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Commission on Biological Macromolecules (2000). *Acta Cryst.* D**56**, 2.
Editorial (1996). *Nature (London)*, **379**, 191.
Editorial (1998*a*). *Nature (London)*, **391**, 617.
Editorial (1998*b*). *Nature Struct. Biol.* **5**, 407–408.
Guex, N. & Peitsch, M. C. (1997). *Electrophoresis*, **18**, 2714–2723.
Hartshorn, M. J. (2002). *J. Comput. Aided Mol. Design.* **16**, 871–881.
Hooft, R. W. W., Vriend, G., Sander, C. & Abola, E. E. (1996). *Nature (London)*, **381**, 272.
Jiang, J., Abola, E. & Sussman, J. L. (1999). *Acta Cryst.* D**55**, 4.
Jones, T. A. & Kjeldgaard, M. (1997). *Methods Enzymol.* **277**, 173–208.
Jones, T. A., Kleywegt, G. J. & Brünger, A. T. (1996). *Nature (London)*, **381**, 18–19.
Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.

Kleywegt, G. J. (1997). *J. Mol. Biol.* **273**, 371–376.
Kleywegt, G. J. (2000). *Acta Cryst.* D**56**, 249–265.
Kleywegt, G. J., Bergfors, T., Senn, H., Le Motte, P., Gsell, B., Shudo, K. & Jones, T. A. (1994). *Structure*, **2**, 1241–1258.
Kleywegt, G. J., Harris, M. R. & Jones, T. A. (2004). To be submitted.
Kleywegt, G. J., Hoier, H. & Jones, T. A. (1996). *Acta Cryst.* D**52**, 858–863.
Kleywegt, G. J. & Jones, T. A. (1996*a*). *Acta Cryst.* D**52**, 826–828.
Kleywegt, G. J. & Jones, T. A. (1996*b*). *Acta Cryst.* D**52**, 829–832.
Kleywegt, G. J. & Jones, T. A. (1996*c*). *Structure*, **4**, 1395–1400.
Koetzle, T. F. (1989). *Nature (London)*, **342**, 114.
Laskowski, R. A., Hutchinson, E. G., Michie, A. D., Wallace, A. C., Jones, M. L. & Thornton, J. M. (1997). *Trends Biochem. Sci.* **22**, 488–490.
Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–291.
Maddox, J. (1989). *Nature (London)*, **341**, 277.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Perrakis, A., Morris, R. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.
Read, R. J. (1990). *Acta Cryst.* A**46**, 900–912.
Reichert, J., Jabs, A., Slickers, P. & Suhnel, J. (2000). *Nucleic Acids Res.* **28**, 246–249.
Vriend, G. (1990). *J. Mol. Graph.* **8**, 52–56.
Wlodawer, A., Davies, D., Petsko, G., Rossmann, M., Olson, A. & Sussman, J. L. (1998). *Science*, **279**, 306–307.