Conference Proceedings
DL-CONF-95-001

Making the Most of your Model

Proceedings of the CCP4 Study Weekend
6-7 January 1995

W N Hunter J M Thornton and S Bailey

June 1995

COUNCIL FOR THE CENTRAL LABORATORY OF THE RESEARCH COUNCILS

© Council for the Central Laboratory of the Research Councils 1995

Enquiries about copyright, reproduction and requests for additional copies of this report should be addressed to:

The Central Laboratory for the Research Councils
Chadwick Library
Daresbury Laboratory
Daresbury
Warrington
Cheshire
WA4 4AD
Tel: 01925 603397 Fax: 01925 603195
E-mail library@dl.ac.uk

ISSN 1358-6254

Neither the Council nor The Laboratory accept any responsibility for loss or damage arising from the use of information contained in any of their reports or in any communication about their tests or investigations.

Making the Most of your Model

Proceedings of the CCP4 Study Weekend
6-7 January 1995

Compiled by:

W N Hunter University of Manchester
J M Thornton University College, London
S Bailey Daresbury Laboratory

CCL
Daresbury Laboratory
1995

CONTENTS

ACKNOWLEDGMENTS

CCP4 would like to thank the EC Human Capital and Mobility Scheme for the provision of funding which allowed 40 young European scientists to attend the 1995 CCP4 meeting.

CCP4 would also like to thank the following companies for their financial contributions to the CCP4 project in the year 1994. This support was an essential contribution to the costs of the meeting.

Abbott Laboratories
Amgen Incorporated
Banyu Pharmaceutical Company Limited
Bristol-Myers Squibb
Chugai Pharmaceuticals
Ciba-Geigy Limited
Daiichi Pharmaceutical Company Limited
Genentech, Incorporated
Genetics Institute
Glaxo UK Ltd.
Green Cross Corporation
Hoescht AG
Hoffmann - La Roche and Co.
Japan Tobacco Incorporated
Pfizer Limited
Pfizer, USA
Pharmacia AB
The Procter and Gamble Company
Sandoz Pharma AG
Syntex (USA) Incorporated
The Wellcome Foundation Limited
Wyeth-Ayrst Laboratories
Zeneca Pharmaceuticals
Zeneca USA

INTRODUCTION

The end product of the crystallographic analysis is a model. From this model we seek to derive an improved understanding of a molecules chemistry and/or how it carries out some biological role. We hope to learn about stability, catalysis and inhibitor action for example and use experimentally determined structures (even NMR models) for homology modelling or to solve other structures. We have also reached a stage where our knowledge about protein structure and function is being directed towards the development of improved methods of structure prediction, new medicines and altered enzymes with improved thermal stability or altered specificity.

At the CCP4 Study Weekend this year we departed from more usual crystallographic topics to concentrate on models. We considered how improved models might might result from improved data collection facilities, how the reliability of the model could be assessed and the current state of predictive methods and fold identification. We also heard how molecular models are being used in a variety of areas. This latter component afforded a more biological view than given at previous meetings.

The meeting was held at Chester College again this year. The facilities available at this venue being suited to the large number of participants. We thank the staff at the college for making us very welcome.

The meeting was organised and supported by the BBSRC Collaborative Computational Project in Protein Crystallography (CCP4) and the EC Human Capital and Mobility Scheme. We thank the invited speakers for sharing their expertise with us and for the contributions to this booklet. We are very grateful to Daresbury Laboratory for providing organisational support, with particular thanks to Val Matthews and Cheryl Stonier and the rest of the SAS team who ensured that the meeting ran to plan.

Bill Hunter
Janet Thornton
Sue Bailey

May 1995

Getting the Best X-ray Intensity Data

Steven E. Ealick

Section of Biochemistry, Molecular and Cell Biology,
Cornell University, Ithaca, NY 14853

Introduction. The success of a structural analysis using X-ray diffraction is ultimately limited by the quality of the experimental intensity data. Whether the goal is structure determination, structure refinement or characterization of structural changes by difference Fourier methods, data quality determines the accuracy of both the phases and the structure factor magnitudes and thus the accuracy of the electron density. The first assessment of data quality usually comes from visual examination of a diffraction image. If the spots are strong, well-resolved and extend to the edge of the detector, one assumes that the data will be good. Barring technical complications during the experiment, ones initial visual impression usually holds true and data that "looks good" usually is good. However, more quantitative measures of data quality are necessary to achieve the maximum benefits from an X-ray diffraction experiment.

Although many factors are associated with data quality, the most commonly used measures are the limit of resolution and the average R-factor between symmetry related reflections (R_{sym}). The limit of resolution is a fundamental property of the crystal but is difficult to quantitate. For example, claims such as "the crystals diffract to at least 2.5 Å resolution" may be based on the observation of a single spot in a diffraction pattern beyond the 2.5 Å limit. In other cases, "2.5 Å resolution data" means that 50% of the data were observed as greater than two times the standard deviation of the measurement. Irrespective of the measure of limit of resolution, the goal of the experiment should be to obtain intensity data at a resolution sufficient to answer the scientific question of interest.

Over the years, the standard for acceptable limit of resolution has risen. Today structures determined at less then 3.0 Å resolution are considered to be low resolution, structures in the 3.0 - 2.2 Å resolution range are considered medium resolution and structures determined beyond 2.2 Å are considered high resolution. The number of structures determined at ultrahigh (1.2 Å or better) resolution is increasing at significant rate.

The most common measure of the accuracy of intensities is R_{sym} defined as:

$$R_{sym} = \frac{\sum_{hkl} \sum_{i} | \bar{I}(hkl) - I(hkl)_i |}{\sum_{hkl} \sum_{i} I(hkl)_i}$$

where $I(hkl)_i$ is the i^{th} measurement for the intensity with indices hkl and $\bar{I}(hkl)$ is the average value for all measurements of that intensity. The R_{sym} value is the most widely used overall indicator of data quality but must be examined carefully to understand its meaning. In general R_{sym} values less than 5% indicate excellent data quality, R_{sym} values in the 5-10% range indicate average data quality, R_{sym} values in the 10-15% range are acceptable but may indicate problems with crystals or experiment and R_{sym} values greater then 15% indicate poor quality data that may not be useful for structural analysis.

In examining the R_{sym} values one should also consider the average redundancy of measurement. High redundancy (more then 5-6 observations per unique reflection) means that the average value of the intensity for a given reflection may be acceptable even when the R_{sym} is high. In addition, one should examine R_{sym} as a function of resolution, intensity and other variables in assessing the quality of data. As a rule of thumb, the R_{sym} should not exceed 25% for the highest resolution shell. Normally the R_{sym} value for the strongest data should be in the 1-2% range. Larger values suggest that the exposure time may be too short. The experimenter should also beware of R_{sym} values that exclude whole classes of reflections such as experimental outliers, weak data or partially recorded reflections. Finally, is should be noted that R_{sym} values are not a normal statistical quantity. As opposed to the goodness-of-fit, the R_{sym} value as described above has no expected valued. Thus a poor R_{sym} value does not indicate whether the fault lies in the sample quality or the experimental design.

Different kinds of experiments require different levels of quality. To determine the structural basis for enzyme mechanism or to design potential drugs based on the three-dimensional structure of an active site requires more detail then determining the arrangement of subunits in the ribosome particle. Determining phases by multiple wavelength anomalous diffraction (MAD) methods requires the measurement of average changes in intensity of a few per cent while multiple isomorphous replacement methods requires measurement of average differences on the order of 15-25 % of the intensity. Protein structure refinement is forgiving of poor data quality because the number of observations is usually small compared to the number of data and because the calculated structure factors contain significant errors themselves. Location of the positions of anomalous scattering atoms requires accurate measurements because the difference in intensity between Bijvoet pairs is small.

Factors affecting the quality of data that can be controlled by the experimenter fall into five main areas: (1) X-ray source, (2) sample, (3) experimental method, (4) data processing and (5) X-ray detector. Choices made by experimenter will determine the quality of data and consequently the limit of its usefulness. Unfortunately, the experimenter is rarely in a position to choose the best of all possible experimental parameters and compromises are usually necessary. For example, synchrotron radiation sources offer considerable benefits over conventional laboratory X-ray sources but require advanced planning and additional expense. Consequently, most experiments are done in the home laboratory. Likewise, one must decide whether to use crystals that diffract to only 2.8 Å resolution or to continue screening crystallization conditions in order to get crystals that diffract to higher resolution. The following sections discuss factors that will aid the experimenter in getting the most out of the X-ray diffraction experiment.

X-ray Sources. Many physical processes produce X-rays and of these three are utilized by most X-ray crystallographers. They are (1) deceleration of an electron by a target material (conventional X-ray sources), (2) electron capture following nuclear decay (radioactive sources) and (3) movement of a charged particle in a magnetic field (synchrotron radiation).

Conventional X-ray sources are of either the sealed tube or rotating anode type. Sealed tube X-ray sources have the advantage of simplicity and low cost. However, production of X-rays by conventional sources is inefficient with about 99.8% of the energy converted to heat for typical anode materials. Therefore, sealed tube X-ray sources are limited to about 2 kW of total power and even less for small focal spots which are most useful for crystals of macromolecules. Consequently, rotating anode generators are the conventional X-ray source most commonly used by macromolecular crystallographers. By spreading the anode material over the surface of a rotating drum, the heat can be more quickly dissipated thus allowing both increased power and smaller focal spots. The disadvantage of rotating anode sources is that they are technically more complicated and require significant routine maintenance.

The most common radioactive source used by crystallographers is Fe^{55}. When Fe^{55} decays to Mn^{55} an electron is captured resulting in the production of an X-ray with a wavelength of about 2.1 Å. Theses sources produce only weak X-ray beams and are generally used for calibration or testing of X-ray instrumentation.

Synchrotron radiation sources have become increasingly popular for macromolecular crystallography (Helliwell, 1993). Ten years ago only a handful of papers reported the use of synchrotron radiation in macromolecular crystallography. Currently, many dozens of structural papers report the use synchrotron radiation each year (Ealick and Walter, 1993). Synchrotron radiation is valuable for macromolecular crystallography because of (1) high X-ray flux, (2) low angular divergence, (3) small source size and (4) continuous spectrum. Also notable are the pulsed time structure and the high polarization of the X-ray beam. These characteristics are valuable for measuring data from crystals with large unit cells, weak diffraction or radiation sensitivity and from small crystals. In addition, the source tunability allows for optimized anomalous scattering and MAD phasing experiments (Hendrickson, 1991). Synchrotron radiation has also proved valuable for measuring accurate high resolution data for structure refinement. Experience at the Cornell High Energy Synchrotron Source (CHESS) suggests that the resolution of diffraction can be improved by as much as 1.0 Å resolution over data collected with conventional X-ray sources.

Figure 1 shows the increase in brilliance of X-rays sources that has been achieved as a function of time. First generation synchrotron radiation sources operated parasitically to particle physics programs through utilization of the radiation created by storage ring bending magnets. The promising results from the initial synchrotron radiation laboratories lead to the construction of dedicated second generation sources which were designed and operated for synchrotron radiation research. The second generation sources also utilized insertion devices called wigglers to increase the total X-ray flux and to shift the critical wavelength to higher energies. Most synchrotron experiments to date have utilized second generation sources.

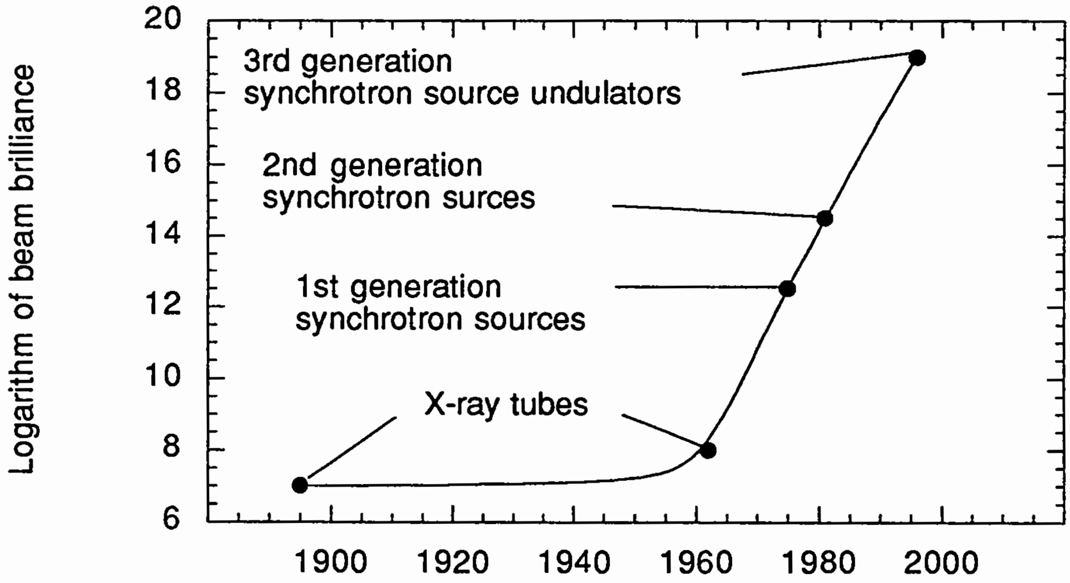


Fig. 1 Plot of logarithm of brilliance versus time for various X-ray sources. Third generation synchrotron sources with undulators offer the highest brilliance.

More recently the first third generation synchrotron source, the European Synchrotron Radiation Source (ESRF), has come on line in Grenoble. Third generation sources are characterized by high energies and small source sizes. These features allow for the operation of undulators which combine wigglers with interference to generate high intensity beams for macromolecular crystallography. Undulator beams have both small source size and low divergence resulting the highest brilliance beams currently available. The Advanced Photon Source is expected to come on line later this year at Argonne National Laboratory. Another third generation source, the 8 GeV Super Photon Ring (SPring-8) is currently under construction in Nishi Harima, Japan.

Synchrotron beam lines require optics to focus the beam and select the energy for the experiment (see Helliwell, 1992 for more details). The most common doubly focused beam line optical arrangements for macromolecular crystallography are (1) single bent crystal for energy selection and horizontal focusing with a mirror for vertical focusing and (2) a double crystal monochromator for energy selection and either a toroidal mirror for double focusing or a vertical focusing mirror combined with bending of the second monochromator crystal for horizontal focusing. The first type of beam line delivers high intensity for monochromatic data collection while the second type of beam line provides rapid tuning and good energy resolution for MAD phasing and optimized anomalous scattering experiments. Synchrotron radiation beam lines are usually maintained by support staffs such that performing the actual experiment requires little special training for the user.

Sample quality. Crystal quality is the most unpredictable variable at the beginning of an X-ray crystallographic investigation. Crystal growth often requires considerable effort and the experimenter tends to use the first crystal form produced. However, additional efforts may lead to improved diffraction quality or a new crystal form with a higher resolution of diffraction. Aside from resolution of diffraction, the mosaic spread is another indicator of crystal quality. Usually mosaic spread is expressed in combination with other factors such as source size, source divergence, energy dispersion and crystal size to give a rocking width, the total angular rotation required to fully record a single reflection. Mosaic spreads for protein crystals are small (usually a few hundredths of a degree) and the rocking width measured with synchrotron radiation is a few tenths of a degree.

Crystals of macromolecules are susceptible to radiation damage and show significant decay in an X-ray beam. At powerful synchrotron sources, the crystal may survive for only a few minutes and fail to yield a complete data set. Multiple crystals can be used to complete the data set but give rise to additional systematic and random errors in the intensities. During the past few years, cryocrystallography has received considerable attention. A technique developed originally by Teng (1990), in which the crystal is suspended in a small loop with a film of mother liquor plus cryoprotectant and then frozen at liquid nitrogen temperatures, has gained wide acceptance. Careful selection of a cryoprotectant usually results in little or no increase in mosaic spread of the crystal. Crystal freezing is now routinely used at synchrotron sources and usually leads to a complete data set from a single sample.

Experimental methods. Several experimental methods have been used to measure X-ray intensities from crystals of macromolecules. They include (1) diffractometry, (2) precession method, (3) rotation/oscillation method, (4) Weissenberg method and (5) Laue method. Diffractometry provides accurate intensities but is time consuming since reflections are measured sequentially. Diffractometers are useful for measuring data from crystals with small unit cells or low resolution data from crystals with larger unit cells. The remaining methods rely upon two dimensional X-ray detectors. The precession method uses sample alignment and camera geometry to generate undistorted diffraction patterns.

The patterns are easy to index but the geometry leads to an efficient coverage of reciprocal space and the need for precise sample alignment.

The rotation/oscillation method is currently the most commonly used data collection method for macromolecular crystallography (Arndt and Wonacott, 1977). Although many variations exist, the basic experiment involves rotation (or oscillation) of the crystal over a fixed angular range while the crystal is exposed to X-rays. The rotation of the crystal is then advanced and the procedure repeated until the data set is considered complete. The limit of rotation is chosen to prevent significant spatial overlap as more and more spots are recorded. The method is mechanically simple and thousands of reflections (or more) can be measured on a single image. The recorded images also lend themselves to indexing and processing by typical computer workstations. The detector may be flat, V-shaped or cylindrical and may be symmetrical with respect to the X-ray beam or offset.

In the rotation method most spots are fully recorded, however, some reflections that cross the Ewald sphere at the beginning or end of the angular range are only partially recorded. The partially recorded reflections can be summed from adjacent data frames or scaled to equivalent fully recorded reflections by dividing the partially recorded intensity by the fraction of intensity that crossed the sphere of diffraction. In some cases, the rotation angular width is chosen to be smaller then the crystal rocking width, sometimes called "phi slicing". Although this strategy makes less efficient use of the detector (i.e. less of the available area is stimulated by X-rays), it has other advantages. In particular, the background for each spot is near the minimum possible value as opposed to wide angle rotations in which background is continuously recorded during the exposure even though the intensity is recorded during only a small part of the range. In addition to increasing the peak to background ratios, this method also allows for three-dimensional profile fitting. Phi slicing is particularly advantageous when the detector readout time is small compared to the exposure time.

The Weissenberg method is in principle another variation of the rotation method in which the detector is translated to minimize spot overlap for prealigned samples (see Stuart and Jones (1993) for a review of macromolecular applications). By translating the detector, spots are less likely to overlap and large rotations (greater than 10 degrees) can be used. The increase in background that results from the larger rotation angle can be compensated by moving a detector, with a large active area, farther away from the sample. Interestingly, comparable efficiencies have been achieved by fixing the detector and intentionally misaligning the sample (Weisgerber and Helliwell, 1993). Under these conditions, surprisingly few overlaps occur even at high resolution. This method would also benefit from detectors with large area located far away from the sample.

The Laue method uses a polychromatic X-ray beam to illuminate a fixed crystal (see Pai, 1992 for a recent review). The Laue method for crystals of macromolecules is only effective when used with a synchrotron radiation source. The amount of data recorded is determined by the crystal orientation and the wavelength range. The Laue method allows substantial amounts of data to be measured in a short period of time (i.e., milliseconds or less) and has been used for time-resolved studies of macromolecules. The primary disadvantages of this method are (1) the coverage of low resolution data is poor and (2) data processing depends on the accurate determination of wavelength dependent correction factors.

Data processing and reduction. A number of data processing programs are now available for determination of integrated intensities. While each program has advantages and disadvantages, two popular programs used at CHESS and other synchrotron sources are DENZO (Otwinowski) and MOSCO (Leslie, 1991). A review of these and other programs for data processing has been published by Finzel (1993). The choice of processing program depends in part on whether data has been collected as wide angle rotations, in a phi slicing mode, or with Weissenberg geometry. Most programs offer profile fitting and background smoothing options. Most programs also provide algorithms to scale and merge data, create a final intensity list and generate tables for assessing data quality. No strong consensus has emerged to support the superiority of one data processing program over another.

X-ray detectors. Dramatic advances have taken place during the past decade in the development of two-dimensional X-ray detectors for macromolecular crystallography (for a brief overview see Pflugrath, 1992). X-ray film was first replaced by multiwire proportional chambers and TV detectors. More recently multiwire detectors have been supplemented by image plate scanners. The most recent development in X-ray detectors involves the use of charge coupled devices (CCD's) to record X-ray diffraction patterns (Gruner, 1994). In general, X-ray detectors can be classified as either photon counting detectors or integrating detectors. Either type of detector works well with conventional X-ray sources, however, high intensity synchrotron sources normally require integrating detectors which are capable of handling the high counting rates. Desirable characteristics for an X-ray detector include high sensitivity, high spatial resolution, large dynamic range, stability over time and a large active area. For synchrotron sources, detectors should also have high counting rates, fast readout and long term stability.

CCD-based X-ray detectors appear to provide the best combination of properties, especially for use at synchrotron radiation sources (Gruner, 1994). These detectors work by imaging an X-ray sensitive phosphor onto a CCD chip with a fiber optic taper. CCD devices and controllers are routinely available in 512 x 512 pixel or 1024 x1024 pixel formats. Larger format CCD's are also available and a 2048 x 2048 pixel detector has been built for use at CHESS. A schematic diagram of one such detector (1024 x1024 format) is shown in Figure 2 (Tate, et al., 1995). The active surface of the detector consists of a $Gd_2O_2S(Tb)$ phosphor and is about 51 x 51 mm^2 giving a pixel size at the phosphor of 50 µm. The phosphor is imaged onto a Thomson TH7896AVRNF CCD using a 2.6:1 fiber optic taper. The CCD chip is cooled by a thermoelectric cooler to minimize dark current noise.

This detector has been in operation at CHESS for over a year and has produced dozens of high quality data sets. An example of a diffraction image measured with the detector is shown in Figure 3. Direct comparison with image plate data measured with a FUJI BAS-2000 off-line scanner suggests that the CCD data is superior in quality, especially at high resolution. A complete data set measured on tetragonal lysozyme yielded an R_{sym} of 2.1% for data to 2.8 Å resolution with an average redundancy of measurement of eight. Data to 2.0 Å resolution showed no increase in R_{sym}, however, the redundancy of measurement at high resolution was much lower. Data measured with this detector has been used for high resolution refinements, molecular replacements structure determinations, difference Fourier analyses of enzyme/ligand complexes, multiple-isomorphous-replacement structure determinations and MAD phasing. Examples of structural results using the 1k CCD detector are shown in Figure 4.

A second CCD-based X-ray detector has been in operation at CHESS for several months. This detector is similar to the first CCD-detector accept that the Reticon CCD has 2048 x 2048 pixels and the fiber optic taper has a 3:1 magnification ratio. This combination results in an active area of 80 x 80 mm^2 and a pixel size of 40 µm. A test data set on tetragonal lysozyme resulted in an R_{sym} value of 2.5% for a highly redundant 2.0 Å resolution data set. Several dozen data sets have now been collected on a variety of projects and structural results are just beginning to emerge.

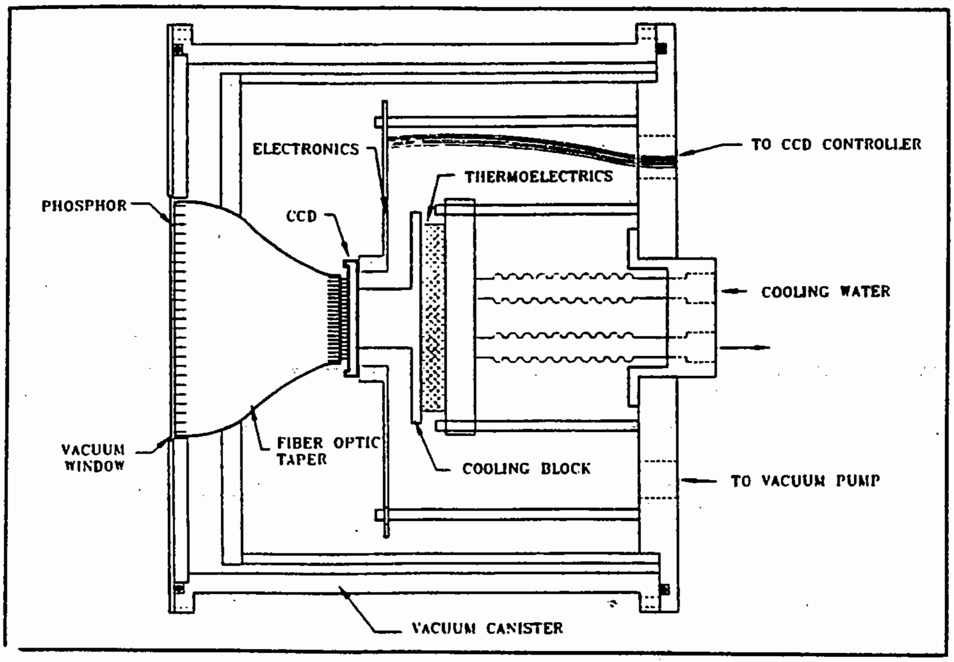


Fig. 2 Schematic diagram of the Princeton 1k x 1k CCD detector installed at CHESS.

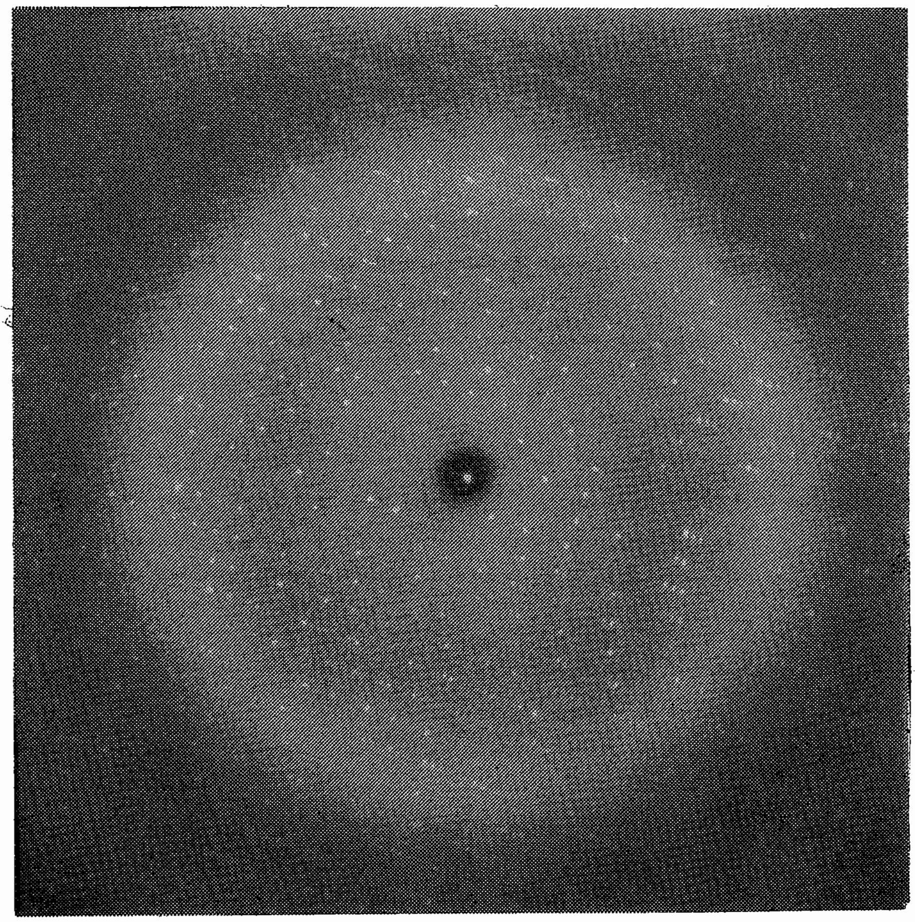


Fig. 3 X-ray diffraction pattern recorded with the Princeton 1k x 1k CCD detector.

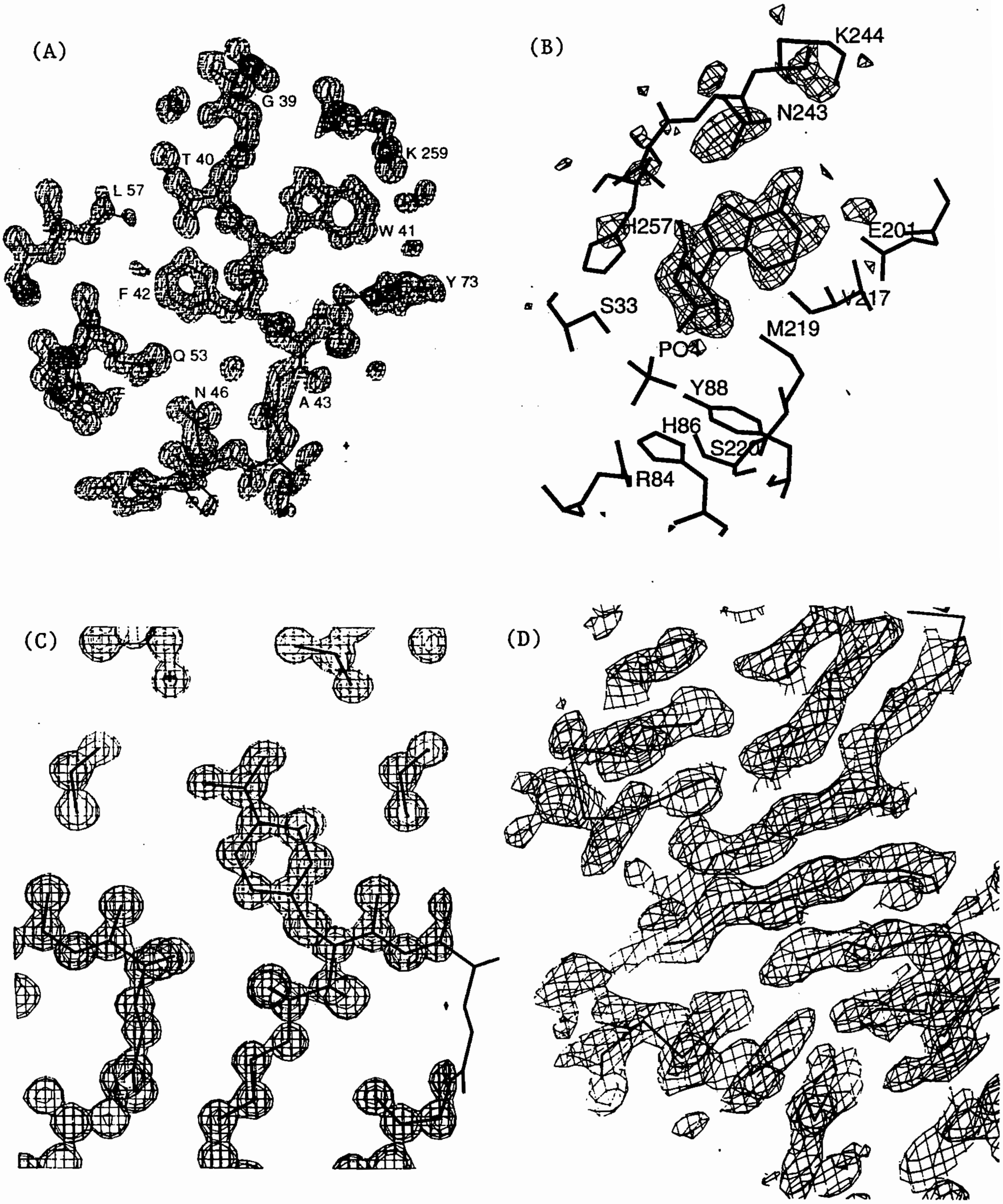


Fig. 4 Examples of results based on data from the 1k x 1x CCD detector at CHESS: (a) 2Fo-Fc map for cellulase calculated at 1.0 Å resolution (Karplus, 1995), (b) difference Fourier map for bovine purine nucleoside phosphorylase plus inhibitor calculated at 1.8 Å resolution (Ealick, 1995), (c) previously unsolvable small molecule determined at 1.0 Å resolution using microcrystals (Clardy, 1995) and (d) segment of the C-terminal receptor domain of the interferon-γ/receptor complex determined at 2.7 Å resolution by MAD phasing (Ealick, 1995).

The main disadvantage of CCD detectors is the small active area. However, the high spatial resolution and low noise allow the detector to be placed close to the sample. At CHESS, the detector size is limiting for less than 10% projects. When spot to spot resolution is a problem, the detector can be moved farther from the sample and offset with respect to the direct beam. Based on experiences at CHESS, the 1k CCD detector can resolve nearly 150 orders of diffraction while the 2k CCD detector can resolve about 200 orders of diffraction. In the future, the size limitations of the 1k and 2k detectors may be overcome by building detectors that consists of several CCD modules arranged in 2x2 or 3x3 mosaics.

Conclusions. Measuring the best data requires selecting the best combination of instrumentation, data collection procedures and data processing methodology. If resources were unlimited one would choose an undulator beam line at a third generation synchrotron source such as the ESRF and utilize short wavelength X-rays to minimize systematic effects such as absorption. Crystals conditions would be found to produce a 1.0 Å resolution limit of diffraction and the data would be collected using the a rotation camera and narrow frame widths. The images would be recorded on a CCD-based X-ray detector (a mosaic detector for a large unit cell) and the data would be processed using three-dimensional profile fitting techniques. The crystals would be frozen at liquid nitrogen temperatures thus allowing integration times long enough to achieve an R_{sym} value of 1-2%.

In reality, the ideal combination of conditions described above is rarely achievable. Nevertheless, one should always keep in mind the nature of the question that the experiment is designed to answer. By so doing, the experimenter can make the appropriate choices, based on the available resources, and get the most from the X-ray diffraction experiment.

References

Arndt, U.W. and Wonacott, A.J. "The Rotation Method in Crystallography", Amsterdam (1977)
Clardy, J.C. private communication (1995)
Ealick, S.E. unpublished results (1995)
Ealick, S.E. and Walter, R.L. Curr Opin Struct Biol, 3 (1993) 725
Finzel, B.C. Curr Opin Struct Biol, 3 (1993) 741
Gruner, S.M. Curr Opin Struct Biol, 4 (1994) 765
Helliwell, J.R. "Macromolecular Crystallography with Synchrotron Radiation", Cambridge University Press (1992)
Hendrickson, W.A. Science, 254 (1991) 51
Karplus, P.A. private communication (1995)
Leslie, A.G.W. In "Crystallographic Computing, Vol 5", edited by Moras, D, Podjarny, A.D. and Thiery, J.C., Oxford University Press (1991) 50
Moffat, K., Szebenyi, D. and Bilderback, D.W. Science, 223 (1984) 1423
Otwinowski, Z. "DENZO, A Program for Automatic Evaluation of Film Densities", New Haven: Yale University
Pai, E.F. Curr Opin Struct Biol, 2 (1992) 821
Pflugrath, J.W. Curr Opin Struct Biol, 2 (1992) 811
Stuart, D.I. and Jones, E.Y. Curr Opin Struct Biol, 3 (1993) 737
Teng, T.Y. J Appl Cryst, 23 (1990) 387
Tate, M.W., Eikenberry, E.F., Barna, S.L., Wall, M.E., Lowrance, J.L. and Gruner, S.M. J Appl Cryst (1995) in press
Weisgerber, S. and Helliwell, J.R. J Chem Soc Faraday Trans, 89 (1993) 2667

Braille for Pugilists

Gerard J. Kleywegt & T. Alwyn Jones,
Department of Molecular Biology,
Biomedical Centre, Uppsala University,
Box 590, S-751 24 Uppsala,
SWEDEN.

Introduction.

Before attempting to make the most out of a model (*e.g.*, looking for structural explanations for biochemical data, designing relevant mutants, modelling as yet unobserved complexes, or constructing potential inhibitors), one has to assess how good the model really is. A thorough, critical analysis of a model requires atomic coordinates and structure factor data, but even from coordinates alone, or even a paper, one can already obtain a fairly good idea of (a) whether or not the model may contain gross errors, and (b) which parts or aspects of the model are credible, and which are to be taken with a grain of salt. At anything worse than atomic resolution, modelling details of protein structures often amounts to an attempt to read Braille while wearing boxing gloves, which means that errors are easily introduced. In this contribution we describe briefly how to minimise errors and artefactual features, and how to assess model quality using a much wider variety of "quality indicators" than is typically done.

What is a good model ?

Quite simply, a good model is one that makes sense in all respects, *i.e.*:

- ***chemical:*** bond lengths, bond angles, *etc.* have reasonable values, chiral carbons have the correct hand, aromatic rings, peptide bonds and conjugated systems are essentially flat.
- ***physical:*** there are no bad contacts, the packing of the molecules in the cell is good, and related atoms (NCS-related, covalently linked, hydrogen-bonded, involved in a salt link) have similar temperature factors.
- ***crystallographic:*** the model adequately explains the experimental data, with as little over-fitting as possible.
- ***protein structural science:*** the protein has some secondary structure, the Ramachandran plot has few if any outliers, peptide oxygens have the correct orientation, the large majority of side chains have a rotamer conformation, salt links and hydrogen bonds make sense, there are no buried charges (except for the odd aspartate), waters and ions are properly coordinated, most if not all peptide bonds are *trans*, *etc.*

- ***statistical:*** the model constitutes the best hypothesis to explain the data, with the minimum degree of over-fitting.
- ***biological:*** the disulfide pairings make sense, and biochemical observations with respect to the effect, of for instance, mutations on the folding or activity can be explained.

Accuracy and precision.

All other things being equal, the best model in a crystallographic sense is the one which has the highest accuracy, and with a precision that matches the information contents of the data. In protein crystallography, "accuracy" is related to $<|\Delta\phi|>$, the average absolute magnitude of the phase errors, and "precision" is related to the level of detail of the model.

The most accurate model, given a particular set of data, is the one that has the lowest value of $<|\Delta\phi|>$, and there are strong indications that the value of $<|\Delta\phi|>$ is highly correlated (correlation coefficient close to +1) with that of the free R-factor, R_{free} [1, 2]. Therefore, the most accurate model is the one with the lowest value of R_{free}.

The level of detail at which one can describe a model is dictated by the information contents (determined by quantity and quality) of the data. For this reason, few people would be tempted to refine anisotropic temperature factors at 2.5Å. On the other hand, many people do refine individual isotropic temperature factors at 3Å or worse resolution, which in most cases amounts to over-fitting of the data. Here, R_{free} can be used to decide if the increased level of detail (*e.g.*, when going from one temperature factor per residue to individual isotropic Bs) is warranted by the data: if refinement of individual Bs leads to a reasonable drop in R_{free}, the new model apparently gives a better description of the data; if it doesn't, refining individual Bs over-fits the data.

Degrees of "wrongness".

Many things can go wrong during building, rebuilding and refinement of a structure, even in the case of molecular-replacement exercises [3]. A few years ago, Brändén and Jones [4] outlined possible degrees of incorrectness of crystallographic models:

- completely wrong model or sub-unit: one in which essentially the entire chain has been traced incorrectly.
- partly wrong main-chain connectivity: usually due to incorrect connections between secondary-structure elements.
- out-of-register error: there is a frame shift for (usually) a small part of the sequence and density.
- locally poor model: either due to bad building or to insufficient data.
- incorrect side-chain conformations.
- incorrect peptide orientations.

Of course there are more possible errors:

- spacegroup error.
- sequence errors: either introduced at the sequencing stage or as a trivial typo.

- ***statistical:*** the model constitutes the best hypothesis to explain the data, with the minimum degree of over-fitting.
- ***biological:*** the disulfide pairings make sense, and biochemical observations with respect to the effect, of for instance, mutations on the folding or activity can be explained.

Accuracy and precision.

All other things being equal, the best model in a crystallographic sense is the one which has the highest accuracy, and with a precision that matches the information contents of the data. In protein crystallography, "accuracy" is related to $<|\Delta\phi|>$, the average absolute magnitude of the phase errors, and "precision" is related to the level of detail of the model.

The most accurate model, given a particular set of data, is the one that has the lowest value of $<|\Delta\phi|>$, and there are strong indications that the value of $<|\Delta\phi|>$ is highly correlated (correlation coefficient close to +1) with that of the free R-factor, R_{free} [1, 2]. Therefore, the most accurate model is the one with the lowest value of R_{free}.

The level of detail at which one can describe a model is dictated by the information contents (determined by quantity and quality) of the data. For this reason, few people would be tempted to refine anisotropic temperature factors at 2.5Å. On the other hand, many people do refine individual isotropic temperature factors at 3Å or worse resolution, which in most cases amounts to over-fitting of the data. Here, R_{free} can be used to decide if the increased level of detail (*e.g.*, when going from one temperature factor per residue to individual isotropic Bs) is warranted by the data: if refinement of individual Bs leads to a reasonable drop in R_{free}, the new model apparently gives a better description of the data; if it doesn't, refining individual Bs over-fits the data.

Degrees of "wrongness".

Many things can go wrong during building, rebuilding and refinement of a structure, even in the case of molecular-replacement exercises [3]. A few years ago, Brändén and Jones [4] outlined possible degrees of incorrectness of crystallographic models:

- completely wrong model or sub-unit: one in which essentially the entire chain has been traced incorrectly.
- partly wrong main-chain connectivity: usually due to incorrect connections between secondary-structure elements.
- out-of-register error: there is a frame shift for (usually) a small part of the sequence and density.
- locally poor model: either due to bad building or to insufficient data.
- incorrect side-chain conformations.
- incorrect peptide orientations.

Of course there are more possible errors:

- spacegroup error.
- sequence errors: either introduced at the sequencing stage or as a trivial typo.

Probably the most common error, however, is:

- over-fitted models: by refining far more parameters than is warranted, the conventional R-factor can be reduced to almost arbitrarily low values, at the expense of a globally worse model. Favourite "tricks" to achieve this include: unrestrained refinement of NCS-related molecules, individual temperature factors at low resolution, fantasised waters, refined occupancies and alternative conformations at medium or low resolution.

Most of these errors can be detected, remedied and (even better) prevented if one uses common sense as well as state-of-the-art methodology (R_{free}, high-temperature Simulated Annealing, databases) and software (*i.e.*, O [5] and X-PLOR [6]).

How NOT to judge a model.

Many journals (Nature, Science, PNAS, to name but a few) are happy with (or insist on) a table with the minimum set of "conventional quality indicators" to convince the readers of the quality of a model: the resolution, the conventional R-factor, the average temperature factor, and the root-mean-square deviation (RMSD) from "ideal values" (often undefined in the paper) of bond lengths and bond angles. These "quality indicators", however, are absolutely unable to discriminate between good and bad models (basically, because they can quite easily be "fudged"). For example, consider the models described in Table I, and try to assess the correctness of both before reading on:

Table I. List of "conventional quality indicators" of two protein models.

Molecule	**"X"**	**"Y"**
Resolution (Å)	3.0	2.9
R-factor	0.214	0.251
RMSD bond lengths (Å)	0.009	0.009
RMSD bond angles (Å)	2.1	1.6
Average temp. factor ($Å^2$)	13.4	49.2

Judging from Table I, and using the conventional ideas as to what constitutes a "well-refined model", model "X" looks quite good, whereas model "Y" has a high R-factor and average temperature factor, indicating that there might be something wrong with it.

In fact, model "Y" is the structure of cellular retinoic-acid-binding protein (CRABP) type I [7]. The R-factor may seem high, but the structure was refined by minimising R_{free} (to get the most accurate model), and by minimising the difference between R and R_{free} (to minimise over-fitting). The temperature factors are high because the quality of the data was less than fantastic (the effective resolution is ~3.2Å), and partly because the structure was refined with strictly constrained two-fold NCS (see the discussion in [7]).

Model “X” is a related protein, CRABP type II [7], which was originally solved at 1.8Å. However, the correct structure was then intentionally traced backwards, and the resulting model was refined using data out to 3Å, to yield model “X” ... [8] Note that this means that the Brändén & Jones 25% R-factor threshold has been broken (this held that a refinement that stalled at an R-factor >0.25 should make “alarm bells ring”).

In other words, the “conventional quality indicators” listed in Table I are not even capable of discriminating between a correct and a backward-traced protein structure ! In the following we shall encounter a number of quality indicators that do a much better job, in particular when they are used in combination (since a good model makes sense in all respects).

Making better models.

The ultimate quality of the structure is determined by the quality of the model building and rebuilding as well as by the refinement protocol. The refinement should always start from a model which has as few assumptions and degrees of freedom as possible, in order to speed up convergence and to limit erroneous adjustments to the model. Initially, this “null-hypothesis” implies:

- there is only protein (no water, no ligand, *etc.*);
- the geometry is near-ideal;
- all NCS-related molecules are identical;
- there is only an overall temperature factor.

The model can then gradually be improved and extended in cycles of rebuilding and refinement. Prior to rebuilding, the model should be checked (“quality control”) on all criteria which are also used to judge the final model: Ramachandran plot, temperature factors, peptide orientations, side-chain conformations, real-space R-factors, geometry, differences between NCS-related molecules, differences with the previous model, *etc. etc.* While rebuilding, the use of databases (for peptide orientations and side-chain conformations, [5, 9]) is essential. At high resolution, it turns out, only ~1-2% of the residues has an unusual peptide orientation, and only ~5-10% of the residues has a non-rotamer side-chain conformation. This means that in the large majority of cases, unusual peptides in early or low-resolution models can be assumed to be wrong (unless the density is extremely well-defined), and that most non-rotamers can safely be replaced by rotamers.

Every refinement cycle, except perhaps the few last ones, should involve high-temperature (4000 K) simulated annealing (SA) [10]. This removes (most of the) model bias and ensures that a large part of conformational space is sampled. It also indicates how “robust” the model is: well-defined parts of the structure do **not** suffer from high-temperature SA, except for the odd side chain.

As the model becomes better and more complete, it can be made more precise (*i.e.*, detailed). For example, the ligand or co-factor can be included in the model, water molecules can be added, the NCS constraints can be replaced by restraints, and temperature factors can be refined for groups of atoms (*e.g.*, two Bs per residue) or perhaps even individual atoms. However, one should realise that each of these steps increases the number of degrees of freedom, and thereby the potential for the refinement program to adjust these parameters in order to model noise and to mask errors in the structure. Undoubtedly atoms have individual (anisotropic) temperature factors, and NCS-related molecules display small differences, but the question one should ask is if the data is of sufficient quality and quantity to actually **model** these phenomena. At anything worse than atomic resolution, R_{free} appears to be the only statistic that can actually tell if an increase in the precision of the model is warranted by the information contents of the data, *i.e.* if it constitutes an improved model for your particular dataset, or if the refinement program has merely used the additional freedom to reduce the conventional R-factor by making the model worse.

What if you don't ?

The refinement and rebuilding protocol outlined above differs rather drastically from the traditional *modus operandi* in the protein crystallographic community. The latter tends to entail unrestrained NCS and individual isotropic temperature factors, **irrespective** of the resolution of the data. An analysis of ~300 low-resolution structures (worse than 2.2Å) reveals that roughly one third has been refined with a data-to-parameter ratio less than one, and an additional one third with a ratio between one and 1.5 [11]. This means that most of these structures suffer from over-fitting, *i.e.* they have been modelled with a level of precision which is not warranted by the data. In the "best" cases, this will have introduced non-existing water molecules, fantasy temperature factors, unrealistic differences between NCS-related molecules and an overall coordinate error of up to 2Å. In the worst cases, the overdose of degrees of freedom will have been used to mask even more serious errors.

A case in point is the structure of chloromuconate cycloisomerase [12] (PDB code 1CHR). This structure was solved in spacegroup I_4 with two-fold NCS. The model was refined against 3Å data without NCS-constraints, with individual temperature factors, with alternative conformations for some residues, and without R_{free}. "Significant differences [...] at the active site" were found between the two NCS-related molecules, and their RMSD was 0.86Å on Cα atoms, and 1.5Å on all atoms. Closer inspection of the model and the data, however, revealed that the actual spacegroup is I_{422}, without NCS [3]. In addition, it was found that a stretch of ~25 residues was out-of-register in the original model [3]. Both errors were masked by the refinement program: since there were ~1.5 times as many degrees of freedom as there were reflections, the wrong model in the wrong spacegroup still had a conventional R-factor of 0.195. Again this shows that the conventional R-factor is rather meaningless at worse than atomic resolution.

The major lessons to be learned from this are:

(1) with conservative models (considerably more observations than degrees of freedom) it is difficult to go wrong, whereas with liberal models (more degrees of freedom than observations) one is almost bound to over-fit and thereby introduce and/or mask errors;

(2) if crystallographically related (and therefore identical) molecules can be "refined" to an RMS positional difference of 1.5Å (provided the resolution is low enough), one has to wonder how different NCS-related molecules really are. We submit that most of the reported differences at medium and low resolution (>2Å) are over-estimates due to over-fitting. Our present "guesstimate" of realistic differences between NCS-related

molecules is ~0.2-0.3Å [11], but more examples of structures with NCS solved at very high resolution are needed. One can only tell if NCS-related molecules are different if one refines them properly, *i.e.* starting with constraints, then refining with restraints (and checking if R_{free} drops), and finally, perhaps, without restraints (and monitoring R_{free}). If one starts off refining without restraints, "observed" differences are nothing but a self-fulfilled prophecy. In some cases, domain movements occur, making RMSD-type comparisons meaningless. But even in those cases one would still expect the "law of conservation of secondary structure" to apply, *i.e.* corresponding residues outside the hinge region(s) should have very similar main-chain dihedral angles.

Judging a paper.

Having to judge a structure merely by reading the paper in which it is described is a situation encountered frequently by readers, editors, referees, co-authors and sometimes even supervisors. Some of the things to check include:

- ***data quality and quantity:*** what was R_{merge}, the completeness, the multiplicity and the strength of the data, both overall **and** in the highest resolution shell ? Have the authors collected a real 1.8Å dataset, or have they mostly collected indices between 2.2 and 1.8Å ?
- ***refinement protocol:*** have the authors carefully designed a refinement protocol suitable to their particular problem ? If R_{free} has not been used throughout the refinement, there is no guarantee that the model is even remotely correct. If no mention is made of grouped temperature factors or NCS-constraints, one may assume that these have not been used and, in particular (but not exclusively) at low resolution, this will have introduced artificial differences between the NCS-related molecules and generally made the model worse than necessary.
- ***model quality:*** what have the authors done to make sure (and convince the reader) that the model is essentially correct ? Does the Ramachandran plot look normal ? Does the temperature-factor plot show regions with consistently high values ? Does the model look like a protein (*i.e.*, in terms of rotamers, peptide orientation, *etc.*) ? If there is NCS, are the molecules similar (RMSD and RMS ΔB on all atoms and on core $C\alpha$ atoms; $\Delta\Phi$ and $\Delta\Psi$) ?

The person who solves the structure has to be absolutely merciless in judging his own model; the supervisor must be supercriticial; even the co-authors should be more critical than the worst nit-picking referee will ever be; the referees should demand to be convinced that the structure is correct; and the editors should start listening to their referees.

Judging coordinates.

When atomic coordinates of the complete model are available, a whole battery of tests can be executed, including those that were not mentioned in the original paper. Table II lists a number of simple checks that can be made with a set of coordinates in hand. The checks have been carried out on a number of models with different degrees of "wrongness":

- ***BACK:*** the backward-traced structure of CRABP type II (model "X" in Table I; [7]);
- ***ASGL:*** the largely incorrect structure of asparaginase/glutaminase [13];
- ***1CHR:*** the original model of chloromuconate cycloisomerase [12], which was solved in the wrong spacegroup and had ~7% of its residues out-of-register [3];
- ***1PMK:*** the structure of a plasminogen kringle domain [14], which is an example of a traditionally refined model at reasonably high resolution;
- two NMR structures (of the glucocorticoid receptor's DNA-binding domain [15]) have been included for comparison, the "final" energy-minimised model (***1GDC***) and model 24 of the ensemble of structures (***2GDA***);
- ***1CBR:*** CRABP I (model "Y" in Table I; [7]) has been included as an example of a conservatively refined low-resolution model.

Finally, two columns have been included which contain our estimates of normal or expected values for the various criteria at high (***NORM***) and low resolution (***LOWR***). The following aspects of the models have been assessed:

1. **Temperature factors:** the model, the average temperature factor and the RMS ΔBs of bonded atoms.
2. **NCS:** RMS positional and temperature-factor differences between NCS-related atoms, differences in the main-chain dihedral angles.
3. **Ramachandran plot:** the percentage of residues in the four types of area as defined by ProCheck [16].
4. **Secondary structure:** the percentage of residues in helices and strands.
5. **Geometry and stereo-chemistry:** some of the properties calculated by ProCheck, the percentage of residues with non-rotamer side chains and unusual peptide orientations, and the overall G-factor.
6. **Directional atomic-contact analysis score:** this measures how (un)usual the ensemble of neighbouring protein atoms is for every group of atoms in the protein [17].

It is clear that none of the traditional quality indicators correlates with the degree of incorrectness of the models. The only exception is the Ramachandran plot, but often this is not included or mentioned in papers in the more prestigious journals. The criteria that correlate best with incorrectness are the percentage of side chains in non-rotamer conformations, the percentage of residues with unusual peptide orientations, and the directional-atomic contact analysis score (DACA). Basically, all three are database methods that provide different ways of probing to what extent a model looks like a real protein. Note that the G-factor calculated by ProCheck can be fudged as well: using the Engh & Huber [18] force field in X-PLOR with not too high a weight for the crystallographic pseudo-energy term virtually guarantees that a structure scores "better than average" in ProCheck (with the exception, perhaps, of the Ramachandran score). If R_{free} had been used in all studies, we are convinced that this statistic would have shown the best correlation with model error, since it is very hard to fudge. On the other hand, some of the structures shown here probably wouldn't have ended up in the literature in the form they did, if R_{free} had been used.

An important conclusion is that an essentially correct model scores well on basically all tests (*i.e.*, makes sense in all respects), a partly incorrect model scores poor on a few tests, and a grossly wrong model scores poor on almost all tests (other than the conventional R-factor and

Table II. Statistics and quality criteria for a number of models with different degrees of incorrectness. See the text for details.

Model	**BACK**	**ASGL**	**1CHR**	**1PMK**	**2GDA**	**1GDC**	**1CBR**	**NORM**	**LOWR**
% Incorrect	100	~80	~7+50	?	?	?	?	0	0
Resolution (Å)	3.0	2.9	3.0	2.25	3-3.5 ?	2.5-3 ?	2.9	1.5-2	>2
Number of residues	137	331	2*370	2*78	72	72	2*136	>50	>50
R	0.214	-	0.195	0.164	-	-	0.251	0.1-0.2	0.2-0.3
Rfree	0.617	-	-	-	-	-	0.320	$<R+R_{merge}?$	
RMSD bond lengths (Å)	0.009	-	0.029	0.015	-	-	0.009	-	<0.015
RMSD bond angles (°)	2.1	-	5.1	-	-	-	1.6	-	<2
Temp.-factor model	B_{iso}	none	B_{iso}	B_{iso}	-	-	grouped	B_{iso}	grouped/none
Average temp. factor ($Å^2$)	13.4	(10)	25.9	22.7	-	-	49.2	5-20?	10-50?
RMS ΔB bonded atoms ($Å^2$)	4.1	-	2.2	2.1	-	-	-	<3?	no B_{iso}
RMSD all NCS atoms (Å) [a]	-	-	1.51	1.17	-	-	0	<0.5	0
RMS ΔB all NCS atoms ($Å^2$) [a]	-	-	5.7	3.7	-	-	0	<5	0
RMSD core Cα atoms (Å) [a]	-	-	0.73	0.71	-	-	0	<0.3	0
RMS ΔB core Cα atoms ($Å^2$) [a]	-	-	4.3	2.3	-	-	0	<3	0
$<|\Delta\Phi|>$ (°) [a]	-	-	23.7	20.9	-	-	0	<5	0
$<|\Delta\Psi|>$ (°) [a]	-	-	23.4	19.3	-	-	0	<5	0
% Residues $|\Delta\Phi| > 10°$ [a]	-	-	60.3	64.1	-	-	0	<5	0
% Residues $|\Delta\Psi| > 10°$ [a]	-	-	60.3	60.3	-	-	0	<5	0
% Core Ramachandran plot areas [b]	42.7	37.0	75.7	64.1	61.9	71.4	81.6	>90	>80
% Additional allowed areas [b]	36.3	31.7	19.4	34.4	30.2	25.4	16.0	5-10	10-20
% Generously allowed areas [b]	12.1	21.0	3.4	0.8	3.2	1.6	1.6	0-3	0-5
% Disallowed areas [b]	8.9	10.3	1.5	0.8	4.8	1.6	0.8	<1	<1

Table II. Continued.

Model	**BACK**	**ASGL**	**1CHR**	**1PMK**	**2GDA**	**1GDC**	**1CBR**	**NORM**	**LOWR**
% Secondary structure [c]	48.9	24.2	62.6	9.6 [h]	50.0	45.8	67.6	50-70	50-70
Omega angle st. dev. (°) [b]	1.6	23.1	7.6	3.2	4.1	4.6	1.5	6?	<2
Zeta angle st. dev. (°) [b]	1.7	5.4	1.0	4.3	2.6	2.6	1.3	4?	<2
Bad contacts per 100 residues [b,e]	13.1	46.2	1.5	37.2	1.4	1.4	1.5	0	<2
H-bond energy st. dev. [b]	0.8	1.6	0.8	1.4	0.7	0.6	0.8	0.5?	<1
% Non-rotamers [c,f]	29.2	29.0	22.4	21.8	13.9	11.1	7.4	5-10	5-10
% Unusual peptide orientations [c,g]	24.1	21.8	4.5	3.8	4.2	1.4	2.2	1-2	1-2
Overall ProCheck G-factor [b]	-0.4	-3.3	-1.3	-1.2	-0.5	-0.5	+0.1	>0	>-0.5
Overall DACA score [d]	-2.6	-2.4	-1.2	-2.0	-2.1	-2.1	-0.4	>-0.5	>-1

[a] - calculated with LSQMAN (GJK & TAJ, unpublished program)

[b] - calculated with ProCheck [16]

[c] - calculated with O [5]

[d] - calculated with What If [17]; this measures how (un)usual the ensemble of neighbouring protein atoms is for every group of atoms in the protein; this may discriminate against DNA-binding proteins (2GDA and 1GDC)

[e] - many hydrogen bonds are flagged as bad contacts

[f] - defined as residues having an RSC-fit value > 1.5 Å

[g] - defined as residues having a pep-flip value > 2.5 Å

[h] - this is a kringle domain (*i.e.*, no α-helices or β-strands)

RMSD values). The same is true, by the way, at the residue level: problematic regions tend to score poor on a number of different criteria (temperature factors, Ramachandran plot, peptide orientation, side-chain conformation, real-space R-factor, *etc.*).

Judging made easy.

If structure factors are available, judging a model becomes a lot easier. First and foremost, it becomes possible to calculate maps which show how good the density really is. Second, using these maps, real-space R-factors [5, 9] can be calculated for each residue. Third, Simulated Annealing omit maps can be used to check poorly defined (or refined) regions. Finally, with the data in hand it is possible to re-do the refinement and to track errors [3], even many years after a structure was first published. Therefore, we strongly encourage the entire protein crystallography community to deposit not only a complete set of atomic coordinates of every solved structure, but also the structure factors with the PDB.

The ultimate test.

A good model can withstand the ultimate test: refinement against high-resolution data. Table III shows two examples of structures from Uppsala which have been solved both at low and at high resolution:

- ***1GUH:*** 2.6Å structure of glutathione S-transferase (GST) A1-1 [19] refined with strict four-fold NCS;
- ***ALEX:*** a non-isomorphous 2Å structure of the same protein refined with restrained two-fold NCS [20];
- ***9RUB:*** a 2.6Å structure of rubisco with two-fold NCS [21];
- ***5RUB:*** the same structure, solved at 1.7Å [22].

Contrary to what one might infer from the table, 1GUH was solved before ALEX, and 9RUB was solved after 5RUB. 1GUH was refined conservatively with strict NCS and grouped temperature factors; 9RUB was refined liberally with no NCS con/restraints and individual temperature factors. Note that the GSTs have very similar values for the majority of the statistics. The GST structures have an RMSD on Cα atoms of 0.47Å, whereas the rubisco's have RMSDs between 0.74 and 1.2Å ! The average $|\Delta\Phi|$ and $|\Delta\Psi|$ is ~9° for the GSTs (rubisco's: 17-20°), and for ~23% of the residues these differences exceed 10° (rubisco's: 53-58%). The average |Δ Cα-Cα-Cα-Cα dihedral| is ~3.7° for the GSTs (rubisco's: 8.0-9.5°), and for 6.9% of the residues this value exceeds 10° (rubisco's: 20-30%). Clearly, the 2.6Å GST model can stand refinement against higher resolution data, whereas the 2.6Å rubisco model has undergone rather large changes which must be due to over-fitting the low-resolution data. Again, the lesson is that conservative refinement minimises the chance of introducing artefacts and errors due to over-fitting, whereas liberal refinement is virtually guaranteed to yield artefacts and errors.

In practice, one often encounters situations in which a structure is first solved at high resolution. This structure is then used to solve the structures of mutants or complexes for which only low-resolution datasets are available. A dangerous mistake is to use the same

Table III. Statistics and quality criteria for two structures from Uppsala that have been solved both at low and high resolution. See the text for details.

Model	**1GUH**	**ALEX**	**9RUB**	**5RUB**
Resolution (Å)	2.6	2.0	2.6	1.7
R/R_{free}	0.229/-	0.196/0.245	0.199/-	0.180/-
Number of residues	4*221	2*221	2*460	2*436
Temp.-factor model	grouped	B_{iso}	B_{iso}	B_{iso}
Average temp. factor ($Å^2$)	35.1	25.5	19.3	29.3
RMS ΔB bonded atoms ($Å^2$)	-	2.7	1.4	1.0
RMSD all NCS atoms (Å) [a]	0	0.57	2.31	1.25
RMS ΔB all NCS atoms ($Å^2$) [a]	0	4.2	7.8	5.3
RMSD core $C\alpha$ atoms (Å) [a]	0	0.09	0.95	0.89
RMS ΔB core $C\alpha$ atoms ($Å^2$) [a]	0	2.1	7.6	5.1
RMS $\Delta\Phi$ (°) [a]	0	3.0	45.9	18.3
% Residues $\lvert\Delta\Phi\rvert > 10°$ [a]	0	2.3	65.3	18.3
RMS $\Delta\Psi$ (°) [a]	0	3.0	45.8	20.0
% Residues $\lvert\Delta\Psi\rvert > 10°$ [a]	0	0.9	67.2	19.2
% Residues $\lvert\Delta$ $C\alpha$-$C\alpha$-$C\alpha\rvert > 5°$ [a]	0	1.4	51.0	12.5
% Residues $\lvert\Delta$ $C\alpha$-$C\alpha$-$C\alpha$-$C\alpha\rvert > 10°$ [a]	0	0.9	39.6	12.5
% Core Ramachandran plot areas [b]	91.9	90.9	74.1	91.0
% Additional allowed areas [b]	8.1	8.4	19.5	8.2
% Generously allowed areas [b]	0	0.8	4.3	0.6
% Disallowed areas [b]	0	0	2.2	0.3
% Secondary structure [c]	69.2	69.0	58.2	63.2
Bad contacts per 100 residues [b,e]	0	0.2	6.3	17.5
% Non-rotamers [c,f]	11.8	10.0	20.0	11.4
% Unusual peptide orientations [c,g]	1.8	2.0	6.8	2.6
Overall ProCheck G-factor [b]	0.0	+0.4	-1.3	-0.4
Overall DACA score [d]	-0.7	-0.6	-1.5	-0.7

[a] - calculated with LSQMAN (GJK & TAJ, unpublished program)
[b] - calculated with ProCheck [16]
[c] - calculated with O [5]
[d] - calculated with What If [17]
[e] - many hydrogen bonds are flagged as bad contacts
[f] - defined as residues having an RSC-fit value > 1.5 Å
[g] - defined as residues having a pep-flip value > 2.5 Å

refinement protocol that was used for the high-resolution refinement for the low-resolution structures (*e.g.*, no NCS restraints, individual temperature factors). The only way in which such structures can be refined properly is by (a) using a conservative refinement strategy, (b) using weak harmonic restraints to keep the atoms near their high-resolution positions unless there is a strong driving force in the data to change them, and (c) monitoring R_{free} from the very start of the refinement. This approach has successfully been applied in the refinement of a complex of *Candida antarctica* lipase B at 2.5Å resolution, starting from a 1.5Å model [23, 24].

Acknowledgments.

This work was supported by the Swedish Natural Science Research Council and Uppsala University. We thank Dr. Alex Wlodawer for providing us with the coordinates of the incorrect model of asparaginase/glutaminase, and Dr. Alex Cameron for the coordinates of the 2Å GST model. We also acknowledge the many fruitful discussions regarding R_{free}, quality control and validation, both with Dr. Axel Brünger (Yale), and with the participants in the European Union Protein Structure Validation Initiative.

References.

[1] A.T. Brünger, Nature 355, 472 (1992).
[2] A.T. Brünger, Acta Cryst. D49, 24 (1993).
[3] G.J. Kleywegt & T.A. Jones, "A more correct crystal structure of chloromuconate cycloisomerase", to be published.
[4] C.I. Brändén & T.A. Jones, Nature 343, 687 (1990).
[5] T.A. Jones, J.Y. Zou, S.W. Cowan, & M. Kjeldgaard, Acta Cryst. A47, 110 (1991).
[6] A.T. Brünger, "X-PLOR: a system for crystallography and NMR", Yale University, New Haven , CT, 1990.
[7] G.J. Kleywegt, T. Bergfors, H. Senn, P. Le Motte, B. Gsell, K. Shudo, & T.A. Jones, Structure 2, 1241 (1994).
[8] G.J. Kleywegt & T.A. Jones, "Refinement of low-resolution structures", to be published.
[9] J.Y. Zou & S.L. Mowbray, Acta Cryst. D50, 237 (1994).
[10] A.T. Brünger & A. Krukowski, Acta Cryst. A46, 585 (1990).
[11] G.J. Kleywegt & T.A. Jones, "Maltreatment of non-crystallographic symmetry", to be published.
[12] H. Hoier, M. Schlömann, A. Hammer, J.P. Glusker, H.L. Carrell, A. Goldman, J.J. Stezowski, & U. Heinemann, Acta Cryst. D50, 75 (1994).
[13] H.L. Ammon, I.T. Weber, A. Wlodawer, R.W. Harrison, G.L. Gilliland, K.C. Murphy, L. Sjölin, & J. Roberts, Proc. Natl. Acad. Sci. USA 263, 150 (1988);
J. Lubkowski, A. Wlodawer, D. Hosset, I.T. Weber, H.L. Ammon, K.C. Murphy, & A.L. Swain, Acta Cryst. D50, 826 (1994).

[14] K. Padmanabhan, T.P. Wu, K.G. Ravichandran, & A. Tulinsky, Prot. Sci. 3, 898 (1994).

[15] H. Baumann, K. Paulsen, H. Kovacs, H. Berglund, A.P.H. Wright, J.A. Gustafsson, & T. Härd, Biochemistry 32, 13463 (1993).

[16] R.A. Laskowski, M.W. MacArthur, D.S. Moss, & J.M. Thornton, J. Appl. Cryst. 26, 283 (1993).

[17] G. Vriend & C. Sander, J. Appl. Cryst. 26, 47 (1993).

[18] R.A. Engh & R. Huber, Acta Cryst. A47, 392 (1991).

[19] I. Sinning, G.J. Kleywegt, S.W. Cowan, P. Reinemer, H.W. Dirr, R. Huber, G.L. Gilliland, R.N. Armstrong, X. Ji, P.G. Board, B. Olin, B. Mannervik, & T.A. Jones, J. Mol. Biol. 232, 192 (1993).

[20] A.D. Cameron, *et al.*, & T.A. Jones, "Structure refinement and analysis of human alpha class glutathione S-transferase A1-1, in the apo form and in complexes with ethacrynic acid and its glutathione conjugate", to be published.

[21] T. Lundqvist & G. Schneider, J. Biol. Chem. 266, 12604 (1991).

[22] G. Schneider, Y. Lindqvist, & T. Lundqvist, J. Mol. Biol. 211, 989 (1990).

[23] G.J. Kleywegt & T.A. Jones, "Good model-building and refinement practice", to be published.

[24] J. Uppenberg, *et al.*, & T.A. Jones, "Crystallographic and molecular dynamics studies of lipase B from *Candida antarctica* reveal a stereo-specificity pocket for secondary alcohols", to be published.

Consistent Stereochemical Dictionaries for Refinement and Model Building

John P. Priestle
Core Drug Discovery Technologies, Pharmaceuticals Research
Ciba-Geigy, Ltd., CH-4002 Basle, Switzerland

In order to build, correct, or refine a protein structure, a list of ideal stereochemical parameters for the amino acids is necessary. These lists are generally called "dictionaries" because the building/refinement program associated with them constantly needs to look up the ideal values for the various stereochemical parameters that define a protein structure. Which parameters need to be included varies with the task at hand. Programs for building and manually correcting a structure have the most modest requirements, generally only bond lengths and bond angles and information about chiral centers and which groups should be planar. Classical restrained least-squares refinement programs usually also include information about van der Waals contacts (although only as a repulsive restraint) and perhaps something about the preferred distributions of torsion angles. Molecular dynamic programs are the most sophisticated of all and attempt to model the entire physiochemical environment of each atom and, in addition to what has been already mentioned, also include attractive and repulsive van der Waals interactions over relatively large distances as well as long range charge interactions, including adding explicit hydrogen atoms for hydrogen bond formation.

The organization of this information for the dictionaries is as varied as the number of programs available employing them, although all must meet some minimum requirement. There must be a way of identifying the stereochemical parameter, the ideal value of this parameter and some weight expressing the confidence one has in this ideal value. From a least-squares point-of-view, one would like to be able to repeatedly measure this ideal value in some unbiased manner. The "ideal" value would then be the average and the appropriate weight would be the reciprocal of the variance of the measurements ($1/\sigma^2$). Another possibility is to express all stereochemical restraints in terms of energy. This has the advantage of connecting the model to the real world and is the only way molecular dynamics can function. For refinement, as opposed to simple stereochemical idealization, one also needs to somehow adjust the stereochemical term to balance with the crystallographic term.

The protein crystallographer has a large selection of refinement programs available to him or her. Among those currently in general use are EREF (Jack & Levitt, 1978), PROLSQ (Hendrickson & Konnert, 1980), TNT (Tronrud *et al.*, 1987), RESTRAIN (Dreissen *et al*, 1989), and the molecular dynamics programs XPLOR (Brünger *et al.*, 1987) and GROMOS (Fujinaga *et al.*, 1989). The most widely used model building programs are FRODO (Jones, 1978) and its many derivatives and O (Jones and Kjeldgaard, 1992). Although the authors go into great detail as to how the stereochemical restraints are applied in their programs, they seldom give much detail about where the ideal values used actually come from. The degree to which one can understand how these dictionaries function and the ease of including new molecular entities varies considerably from program to program. Some examples of how these dictionaries function are given below:

FRODO / O

Although different programs employing different building algorithms and philosophies, both FRODO and O use the same ideal values and the Hermans and Ferro (1971) algorithm for protein building and stereochemical idealization. The entry for alanine is given in Table 1.

```
ALA          10      *****
N    1-1 2 0 0 1.32  114  180 24.13 637.9-0.204
CA   2 1 3 4 1 1.47  123  180 19.64 957.0 0.058
CB   1 2 0 0 0 1.53  110 -120 19.64 957.0-0.120
C'   2 2 5 6 2 1.53  110  180 23.65 897.6 0.318
O    0 4 0 0 0 1.24  121  180 23.25 421.5-0.422
```

Table 1. Entry for alanine from the dictionary for the model building programs FRODO and O (hydrogen atoms removed).

The first column after the atom names is not used. The next three columns describe branching to and from that atom, while the fifth column is a torsion angle flag. The sixth column is bond lengths (to the previous atom) and the seventh column is bond angles (to the previous two atoms). The eighth column is default torsion angles (proper or improper) through that atom. The other columns contain information about van der Waals parameters, partial charges, etc. This system has the advantage that it allows the *de novo* construction of a protein structure either from a point or as an extension of a preexisting chain. It also has the advantage that the main chain atoms of the individual amino acids can be separately defined. Its major disadvantage is its complex bookkeeping, making addition of new chemical entities a heroic effort. The weights associated with the parameters are all the same for the parameter type (bond length, bond angle or torsion angle) and are set within the program itself.

PROTIN / PROLSQ

This program suite, along with EREF, was one of the first restrained least-squares refinement programs for proteins and still continues to be extremely popular. The program PROTIN sets up the stereochemical restraints, while PROLSQ is the actual refinement program. The section for alanine from PROTIN's dictionary is shown in Table 2.

```
                                      1   -1        MAIN
   1.20134   0.84658   0.        2         1    1           N
   0.        0.        0.        1         2    2           CA
  -1.25029   0.88107   0.        1         3    3           C
  -2.18525   0.66029  -0.78409   3         4    4           O
                                      1    1        ALA A
  +0.02022  -0.92681   1.20938   1         5    5           CB

ALA           1    8
  1  2 1 1  2  3 1 1  3  4 1 1  1  3 2 2  2  4 2 2  5  2 1 3
  5  3 2 4  5  1 2 4
ALA           1    2
  1  4 1  4  5 1
ALA   1  0    3    1    2    3    1    2
```

Table 2. Entry for alanine from the dictionary for the refinement program PROLSQ.

The PROTIN dictionary gives coordinates for an ideal residue with its Cα at the origin. These coordinates are drawn from a single, very high resolution, well refined small molecule crystal structure. All residues are broken into main chain and side chain components. There then follows a table where bonds are described. Each bond has four parameters, the two atoms which define it, its type (1-2, 1-3 or 1-4 bond) and whether it is a main chain, side chain, or mixed bond. Bond angles are expressed as 1-3 bond distances while 1-4 bond distances are used to restrain fixed torsion angles like those found in aromatic rings and the peptide bond. There are similar tables for planar groups (alanine has none, since the peptide bond is defined in MAIN), chiral centers (all Cα are defined as the same), possible intraresidue van der Waals contacts and torsion angles. The bookkeeping for the PROTIN dictionary is less complicated than for FRODO/O. Some disadvantages of the PROTIN system are that all amino acid main chains are considered identical and all restraints of the same type have the same weight and are set within the program PROLSQ.

TNT

One the main considerations of the authors of TNT was that it should be fast and that it should be easy to define new chemical entities in the stereochemical dictionary. Since TNT uses Fast Fourier Transforms (FFTs) for calculating structure factors and their derivatives (like EREF), it was much faster than PROLSQ which did a full trigonometric expansion. FFT versions of PROLSQ have since appeared (e.g. Finzel *et al.* 1987), so that the execution time difference is minor. However, TNT does have a format for its stereochemical dictionary that is hard to beat for simplicity and straightforwardness (Table 3).

```
GEOMETRY PEPTIDE   BOND     1.45    0.02   N, CA
GEOMETRY PEPTIDE   BOND     1.52    0.02   CA, C
GEOMETRY PEPTIDE   BOND     1.33    0.02   C, +N
GEOMETRY PEPTIDE   BOND     1.23    0.02   C, O
GEOMETRY PEPTIDE   ANGLE    112.0    3     N, CA, C
GEOMETRY PEPTIDE   ANGLE    121.1    3     CA, C, O
GEOMETRY PEPTIDE   ANGLE    115.6    3     CA, C, +N
GEOMETRY PEPTIDE   ANGLE    123.2    3     O, C, +N
GEOMETRY PEPTIDE   ANGLE    121.9    3     C, +N, +CA
GEOMETRY PEPTIDE   PLANE    5       0.02   C, CA, O, +N, +CA
GEOMETRY PEPTIDE   TORS     2160     30    N, CA, C, +N
GEOMETRY PEPTIDE   TORS     2180     10    CA, C, +N, +CA
GEOMETRY PEPTIDE   TORS     3060     20    C, +N, +CA, +C
GEOMETRY ALA       BOND     1.52    0.02   CA, CB
GEOMETRY ALA       ANGLE    110.9    3     N, CA, CB
GEOMETRY ALA       ANGLE    111.0    3     C, CA, CB
GEOMETRY ALA       CHIRAL   1        1     CA, N, CB, C
```

Table 3. Entry for alanine from the dictionary for the refinement program TNT.

Each line consists of the keyword "GEOMETRY", the residue name, the type of restraint, its value, its standard deviation (square root of the variance) and a list of atoms that define the restraint. It suffers from the same restriction as PROLSQ in that the main chain of all residues are treated as identical. It does have the extra advantage that each individual restraint has its own weight, although as seen above, this is not fully taken advantage of in practice.

X-PLOR

The dictionary for X-PLOR works differently from the previous ones in that the atoms of each residue are assigned atom types (in a topology file) and the stereochemical parameters are then listed by atom type (in a parameter file). The topology for alanine and a sample of the parameter file for X-PLOR are given in Tables 4 and 5, respectively.

```
RESIdue ALA
 GROUp
  ATOM N     TYPE=NH1    CHARge=-0.35    END
  ATOM H     TYPE=H      CHARge= 0.25    END
  ATOM CA    TYPE=CH1E   CHARge= 0.10    END
  ATOM CB    TYPE=CH3E   CHARge= 0.00    END
  ATOM C     TYPE=C      CHARge= 0.55    END
  ATOM O     TYPE=O      CHARge=-0.55    END

 BOND N     CA
 BOND CA    C
 BOND C     O
 BOND N     H
 BOND CA    CB

 IMPRoper  CA      N     C  CB  !tetrahedral CA

 DONOr H     N
 ACCEptor O C

END {ALA}
```

Table 4. Topology entry for alanine from the dictionary for the molecular dynamics program X-PLOR. Angle restraints do not need to be explicitly given because the program can generate them from the bonded atom list.

```
bond C     CH1E    405.0  1.52!  EXCEPT WHERE NOTED.
bond C     O       580.0  1.23
bond CH1E  CH3E    225.0  1.52
bond CH1E  NH1     422.0  1.45
bond H     NH1     405.0  0.98!  GELIN AND IR STRETCH 3200 CM 1

angle CH1E C    O       85.0 121.5
angle C    CH1E CH3E    70.0 106.5
angle C    CH1E NH1     45.0 111.6
angle CH1E CH1E CH3E    45.0 111.0
```

Table 5. Some parameters from the dictionary for the molecular dynamics program X-PLOR.

Being a molecular dynamics program, X-PLOR requires more parameters to define the atomic environment than do classical least-squares programs like PROLSQ and TNT. In particular, atomic masses (not shown in Table 4) and partial charges are assigned in the topology file. Explicit hydrogen atoms on heteroatoms need to be included (their positions are calculated by the program). The van der Waals interactions are described more fully with a 6-12 $(A/r^{12} - B/r^{6})$ Lennard-Jones potential. The advantage of first defining atom type and then their parameters is that a lot of duplicated definitions can be avoided. For example, the bond length between two methylene groups needs only be defined once. Each restraint has its own weight, here expressed as a force constant (first column after the defining atoms) followed by the ideal (equilibrium) value. One has to be careful, however, that the atom types are properly defined. For example, an aromatic carbon in a six-membered ring will have different parameters than those in a five-membered ring. Each amino acid is defined in full so that subtle differences in the main chain, e.g. for glycine and proline, can be taken account of. It should be noted, that the force constants are not related to the accuracy of the parameter measurement, as in classical least-squares, but to the actual energies of the physical atomic model.

Despite the quite different ways of introducing stereochemical restraints, all these programs use parameters which describe the same systems, the amino acids. One might assume that their values are quite similar, at least within the presumed accuracy of the parameters. In fact,

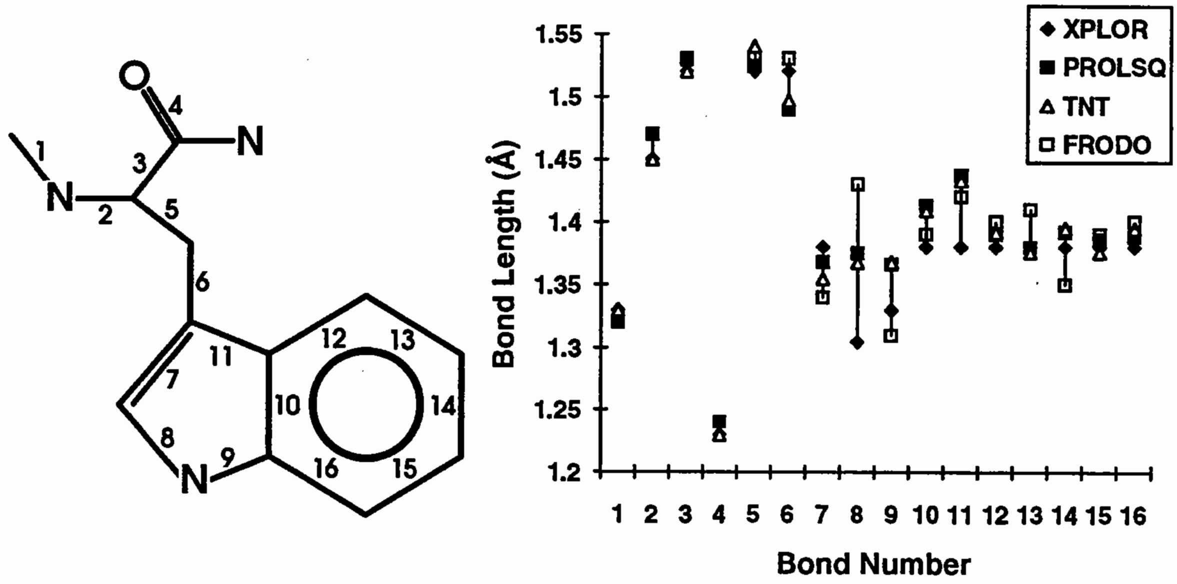


Figure 1. Analysis of the "ideal" bond lengths for tryptophan from various refinement and model building programs.

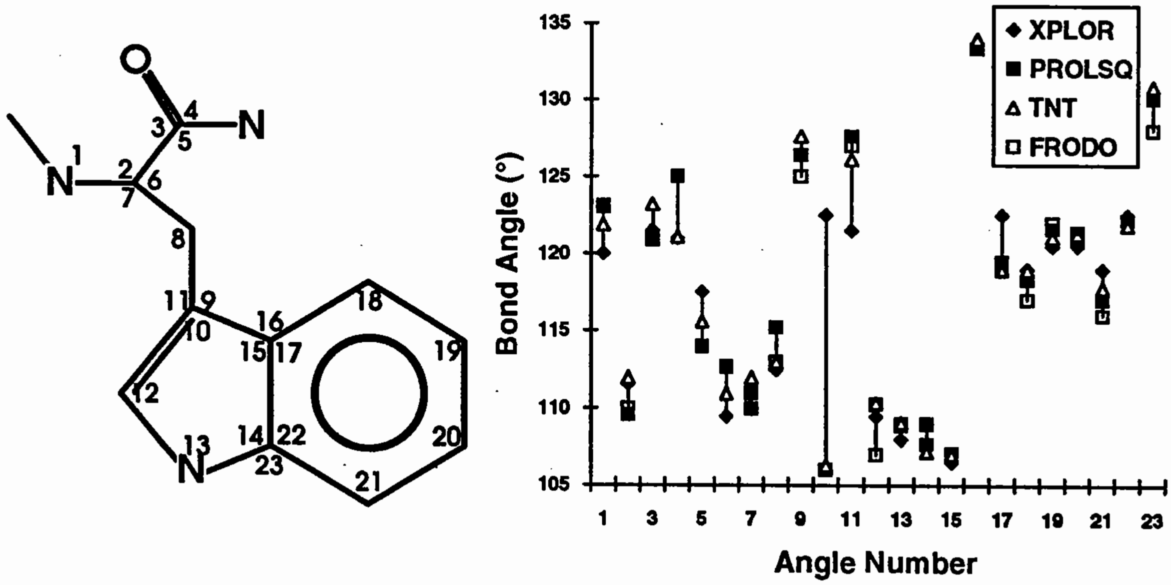


Figure 2. Analysis of the "ideal" bond angles for tryptophan from various refinement and model building programs. The large deviation for X-PLOR for angle 10 results from its being defined by the same atom types as angles 17 & 22 (C-C-CR1E).

they are not. Figs. 1 and 2 show an analysis of the bond lengths and angles associated with tryptophan (admittedly an extreme case) from the programs in use in our laboratory. The differences between the various dictionaries are often much larger than the supposed accuracy of the values (typically $\sigma \sim 0.02$Å for bond lengths and ~ 3° for bond angles).

What's interesting to note is that the scale of disagreement between the various programs varies considerably among the different residues, with heteroatom-containing side chains causing more problems than hydrophobic ones (Fig. 3). Even more interesting is that the different programs seem to have problems with different amino acids (Fig. 4).

The question naturally arises whether these inconsistencies affect the quality of the final model of the protein structure or will the X-ray contribution be able to overcome this dictionary bias. This was investigated by Laskowski *et al.* (1993) who were actually looking at the effect of resolution on the final stereochemical parameters, their hypothesis being that higher resolution should mean more diffraction data and less dependence on stereochemical restraints. They found no such correlation, but they also examined the use of different refinement programs and their associated stereochemical dictionaries and asked whether these differences left an imprint on the final, fully refined structure. By a rather crude analysis of bond lengths and bond angles, they were able to correctly identify the refinement package used for a particular protein structure 95% of the time(!), implying that the program and its dictionary does indeed leave a strong bias on the final structure. This also suggests that if you go from one program to another, one program's idea of idealization is another program's idea

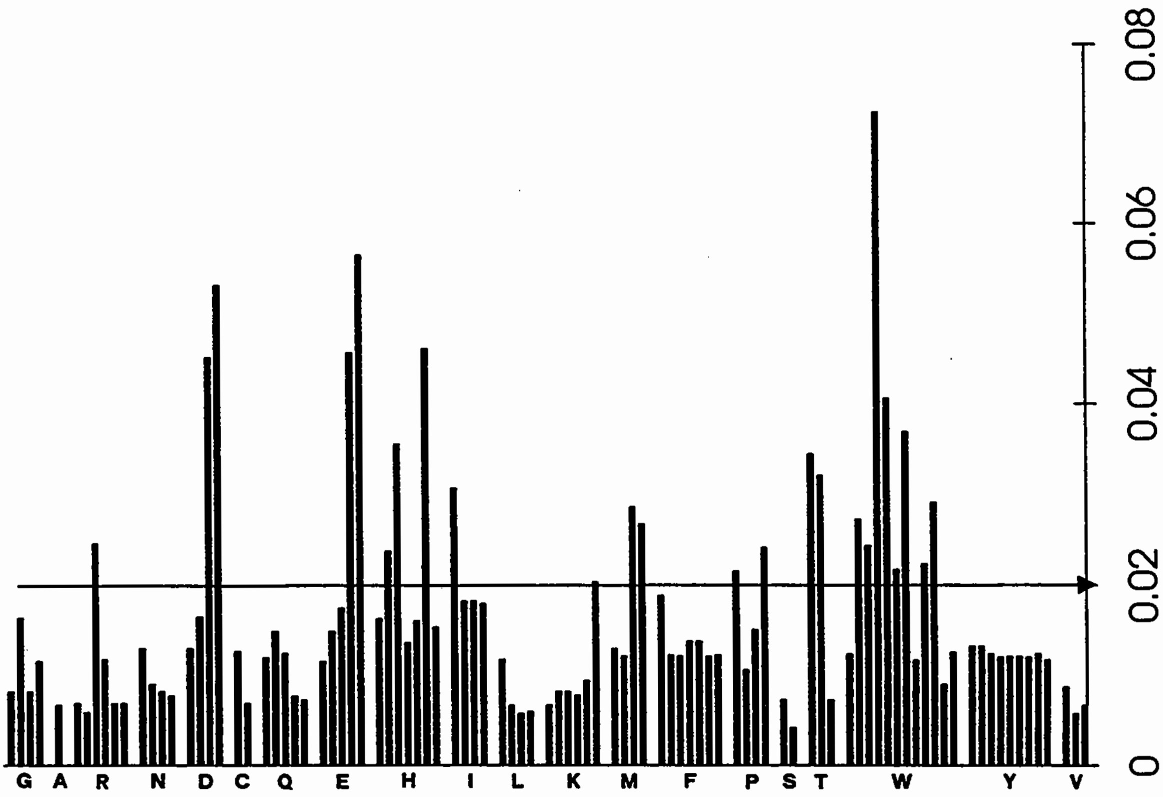


Figure 3. RMS deviations of "ideal" bond lengths for the side chains of the amino acids from various refinement and model building programs. The bond lengths for glycine are between the main chain atoms. The arrow line through 0.02Å is the estimated accuracy of bond lengths.

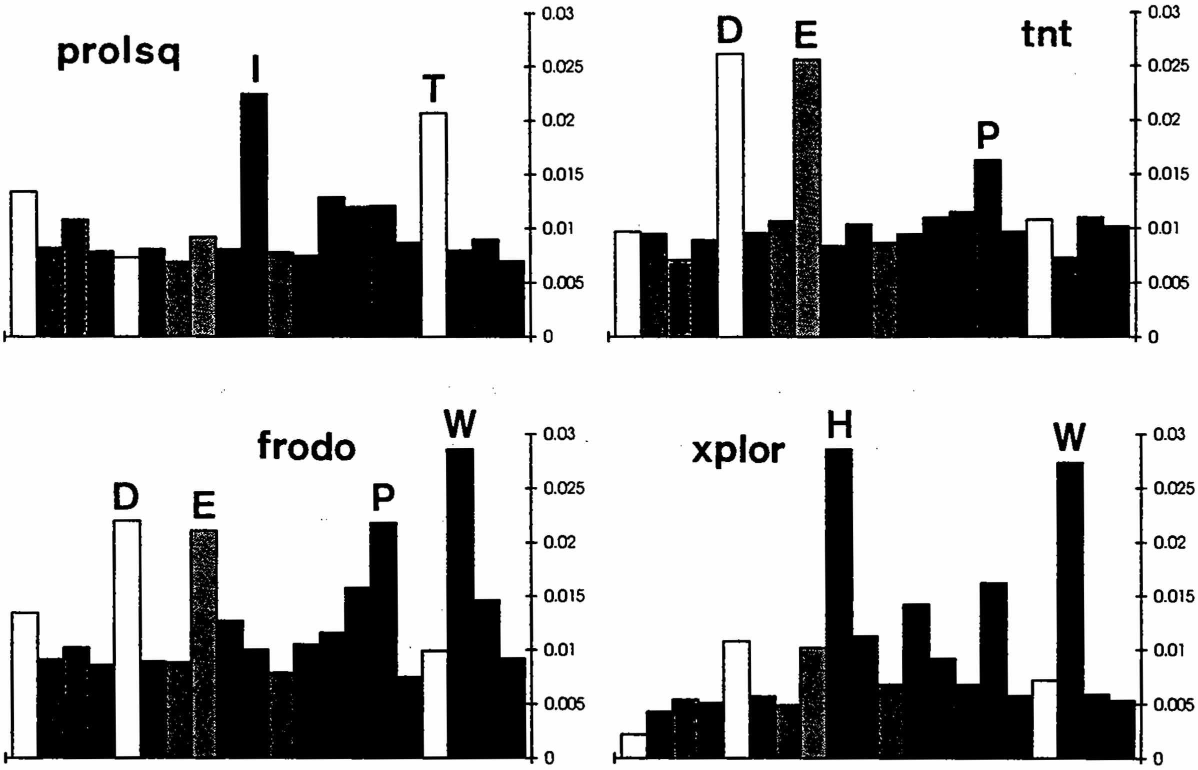


Figure 4. RMS deviations of the amino acid bond lengths compared to the most accurately determined values (Engh & Huber, 1991). Some deviations are explainable, e.g., in TNT aspartic and glutamic acids are assumed to be in the protonated form, rather than the more predominant charged form and in X-PLOR, residues with 5-membered aromatic rings use carbon atoms types for 6-membered rings.

of distortion. Clearly, a set of standard stereochemical parameters that is consistent across all programs is highly desirable.

The need for a set of accurate standard stereochemical parameters was recognized by Engh and Huber (1991) who systematically investigated amino acids and related structures in the Cambridge Structural Database of small molecule X-ray structures to get a set of best bond distance and bond angle parameters for the amino acids, and almost equally important, an accurate estimation of their variance. Their parameters were written up and distributed as topology and parameter files for use with X-PLOR (topology file "tophcsdx.pro" and parameter file "parhcsdx.pro"). They converted the standard deviations of the measurements to force constants by finding those force constants which, in the absence of other forces, would result in a distribution about the mean dictated by the standard deviations of the measurements. Two test proteins using these new parameters demonstrated a small decrease in the R-factor (0.1-0.3%) and a global decrease in the energy of the system, despite the fact that the new parameters have much larger force constants (~3x larger for the bond lengths and ~7x larger for bond angles).

The stereochemical dictionaries for some of the other refinement and model building programs (PROLSQ, TNT, FRODO and O) have recently been rewritten to be consistent with the Engh & Huber X-PLOR parameters (Priestle, 1994). Some dictionaries were easier to modify than others. The dictionary for TNT was the easiest to change, which should be satisfying to the program's authors, since ease of modification/addition to the dictionary was one of the driving forces behind that program's development. The FRODO and O dictionaries were much more difficult, foremost because of the complex bookkeeping. A second problem was determining accurate pseudo-torsion angles at branch points. The most difficult and error-prone dictionary was that for PROLSQ. Because the parameters are not given directly, but calculated from coordinates, this meant that the parameters of Engh & Huber had to be first converted to "ideal" coordinates, from which the "ideal" parameters can then be derived. A second problem was monitoring the angle parameters which are not expressed as angles, but rather as 1-3 bond distances. A dictionary for PROLSQ based on the Engh & Huber parameters, but even more self-consistent than the one presented here, has also been recently developed by Lamzin *et al.* (1994).

In comparing the agreement between the various programs the root mean square deviations between the various "ideal" amino acid bond lengths have been reduced from 0.031Å to 0.003Å for bond distances and from about 1.9° to 0.3° for bond angles. Perfect agreement is not possible because of the different ways some groups are handled by the different programs. For example, PROLSQ and TNT treat the main chain of all amino acids as being identical, while other programs allow for the subtle differences in, for example, proline and glycine. Likewise, FRODO and O do not take special considerations for the amino and carboxyl terminal groups. In any case, a significant improvement in the agreement between the various programs has been effected. The updated stereochemical dictionaries are available from the CCP4 secretary (Daresbury Laboratories, England) or they may be obtained directly from the author (E-mail address: PRIESTLE@FMI.CH).

REFERENCES

Brünger, A.T., Kuriyan, J. and Karplus, M. Science *235* (1987) 460

Dreissen, H., Haneef, M.I.J., Harris, G.W., Howlin, B., Khan, G. and Moss, D.S. J. Appl. Cryst. *22* (1989) 510

Engh, R.A. and Huber, R. Acta Cryst. *A47*(1991) 392

Finzel, B.C. J. Appl. Cryst. *20* (1987) 53

Fujinaga, M., Gros, P. and van Gunstern, W.F. J. Appl. Cryst. 22 (1989) 1

Hendrickson, W.A. and Konnert, J.H. in *Computing in Crystallography* (Diamond, R., Ramaseshan, S. and Venkatesan, K., eds.) Indian Acad. Sci., Bangalore (1980) p. 13.01

Hermans, J. Jr. and Ferro, D. Biopolymers *10* (1971) 1121

Jack, A. and Levitt, M. Acta Cryst. *A34* (1978) 931

Jones, T.A. J. Appl. Cryst. *11* (1978) 268

Jones, T.A. and Kjeldgaard, M. (1992) O - The Manual, Uppsala, Sweden.

Lamzin, V.S., Dauter, Z. and Wilson, K.S. J. Appl. Cryst. (1994) *In Press*

Laskowski, R.A., Moss, D.S. and Thornton, J.M. J. Mol. Biol. *213* (1993) 1049

Priestle, J.P. Structure 2 (1994) 911

Tronrud, D.E., Ten Eyck, L.F. and Matthews, B.W. Acta Cryst. *A43* (1987) 489

IMPLICATIONS OF ATOMIC RESOLUTION

Victor S. Lamzin, Jozef Sevcik*, Zbigniew Dauter and Keith S. Wilson

European Molecular Biology Laboratory (EMBL), c/o DESY,
Notkestraße 85, D-22603 Hamburg, Germany

*** Institute of Molecular Biology, Slovak Academy of Sciences,**
Dubravska cesta, 842 51 Bratislava, Slovak Republic

ABSTRACT

Crystals of RNase Sa diffract to atomic resolution at room temperature. It is important to record the best possible data. Using synchrotron radiation and an imaging plate scanner X-ray data have been collected for native enzyme and a nucleotide complex. Models were refined using SHELXL–93 with anisotropic atomic temperature factors in an essentially automatic manner to R factors of 10.6 % at 1.20 Å (native RNase) and 10.9 % at 1.15 Å (complex with guanosine-2'-monophosphate). The r.m.s. error in the coordinates is 0.05 Å. It is substantially less than that obtained for structures refined isotropically. Analysis of peptide planarity and torsion angles is presented. Some bond lengths are found to differ from the targets widely used for protein structure refinement. The solvent structure was modelled using objective criteria and is compared for models refined using different programs with isotropic and anisotropic description of atomic thermal motion.

INTRODUCTION

For crystal structure determination of small molecules X-ray data typically extend to 1.0 Å or to the edge of the CuKα radiation sphere. This is sufficient to define a good least-squares minimum and to refine atomic parameters against these data alone. The situation is different for proteins. Their crystal unit cells contain a much larger number of atoms. This results in substantially weaker diffraction. In addition a protein crystal contains about 50 % solvent. The limited number of X-ray data is a severe complication in protein crystallography for both structure solution and refinement.

The maximum number of independent reflections which can be collected from a crystal depends on the crystal packing and resolution (Blundell & Johnson, 1976). For a typical protein crystal at 2.5 Å resolution there are 4.5 reflections per atom, just sufficient to refine 3 positional and 1 thermal atomic parameters, Figures 1 and 2. Well ordered solvent can be identified. A resolution lower than 2.5 Å is really low and comprehensive refinement is impossible. At high resolution, 2.0 Å or better, there are 9 or more observations per atom which allows the satisfactory refinement of 4 atomic parameters and for the isotropic model gives a ratio observations/parameters of more than 2. Hydrogens can be introduced and some double conformations become visible in the density map. If the resolution approaches 1.5 Å the number of

Resolution (Å)	N / atom	N / parameter Isotropic	N / parameter Anisotropic
1.0	70	17.6	7.8
1.2	41	10.2	4.5
1.5	21	5.2	2.3
2.0	9	2.2	1.0
2.5	4.5	1.1	not enough
3.0	2.6	not enough	not enough

Figure 1. Expected number of independent reflections (N) for a protein crystal with V_m of 2.4 Å^3/Da at different resolution.

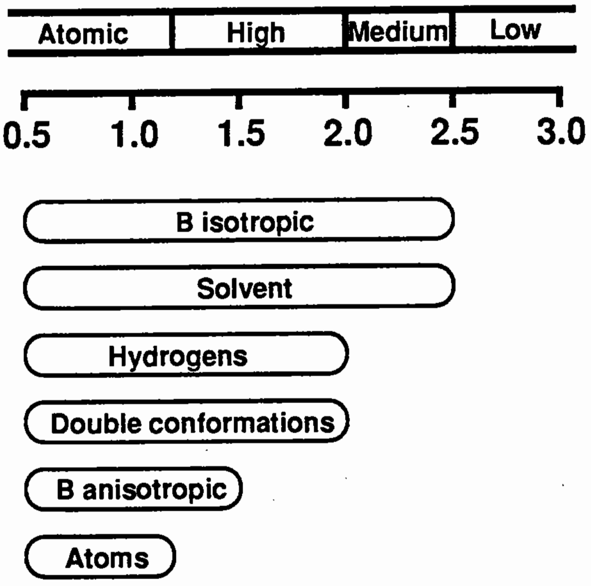


Figure 2. Resolution for protein crystals.

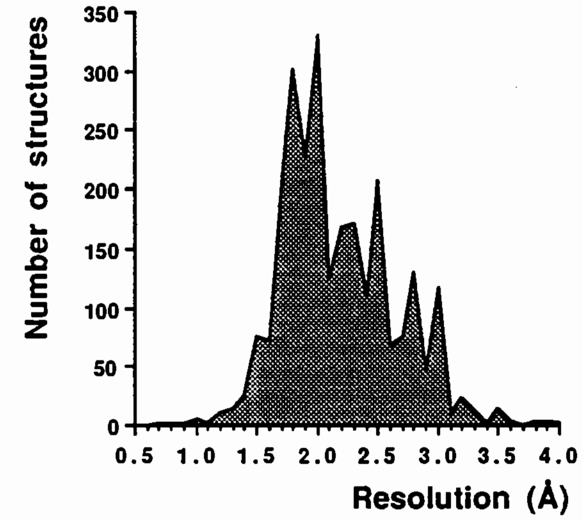


Figure 3. Resolution quoted for protein crystal structures in PDB (Jan. 1995)

observations is enough to refine atomic temperature factors anisotropically. At a resolution of 1.2 Å or better, the density map clearly reveals atomic features. However both protein and small molecule crystals showing diffraction to atomic resolution usually have low solvent content and the number of independent reflections shown in Figure 1 can be less by up to a factor of 2.

Although the limited number of observations for protein crystals even at a resolution less than atomic exceeds the number of parameters, it is not enough to define them precisely. Additional information needs to be introduced as restraints. Even at atomic resolution it is not possible from the X-ray terms alone to define atoms which are poorly ordered or present in multiple conformations, as is also found for small molecule structures and weak restraints are required. The X-ray data must be complemented by stereochemical parameters based on a chemical knowledge of organic structures. The most recent set was derived from the small molecule

Cambridge Data Base (Engh & Huber, 1991). One aim of the present series of atomic resolution protein structures at EMBL Hamburg is the derivation of a library of protein stereochemistry based on real proteins. Even though about 50 % of the crystal protein structures in the Protein Data Bank (Bernstein *et al.*, 1977) are quoted as high resolution, few of them have been refined at truly atomic resolution, Figure 3.

X-RAY DATA

X-ray data collection is the most important step in structure determination after crystals have been grown. Well measured data to the highest tractable resolution will make all subsequent steps in structure refinement and analysis easier. Fortunately data collection for protein crystals is no longer a rate limiting step and is technically easy due to the availability of synchrotron radiation and sensitive two-dimensional detectors.

As an example we have chosen ribonuclease from *Streptomyces aureofaciens.* It highly specifically hydrolyses the phosphodiester bonds of RNA at the 3'-side of guanosine nucleotides and belongs to the prokaryotic subgroup of the microbial ribonuclease family. The molecule consists of 96 amino acid residues (Shlyapnikov *et al.*, 1986). Several crystal structures of native RNase Sa and its complexes were previously determined up to 1.7 Å resolution (Sevcik *et al.*, 1993). It was clear that higher resolution data could be collected.

RNase Sa native crystallises in space group $P2_12_12_1$ (Sevcik *et al.*, 1993). There are two independent molecules in the asymmetric unit, referred to throughout as molecules A and B. The complex was prepared by diffusion of 2'-GMP into native crystals. Data were collected at room temperature from a single crystal of native enzyme (1.20 Å resolution, Rmerge(I) 3.9 %) and two crystals of the complex (1.15 Å resolution, Rmerge(I) 6.6 %) on the EMBL X11 beamline at the DORIS storage ring, DESY, Hamburg with a MAR research imaging plate scanner. Data are essentially complete at low resolution and about 90 % complete overall. The MOSFLM (Leslie, 1992) and DENZO (Otwinowski, 1993) packages were used. Radiation with highly accurately calibrated wavelength was used to determine accurate unit cell parameters (a = 64.73 (4), b = 78.56 (7), c = 38.99 (5) Å) with an accuracy of about 0.1 %.

REFINEMENT AT ATOMIC RESOLUTION

Each refinement cycle consisted of least-squares minimisation, Fourier map calculation and automated updating of solvent, Figure 4. The 1GMQ set (waters omitted) was used as starting model. Refinement was carried out against 95 % of the data. The remaining randomly excluded 5 % were used to calculate the Rfree factor (Brünger, 1993). Atomic positions and isotropic temperature factors were refined by restrained least-squares minimisation against F's using the CCP4 (1994) version of PROLSQ (Konnert & Hendrickson, 1980). Anisotropic refinement against F^2's used SHELXL-93 (Sheldrick, 1993). Hydrogen atoms were introduced on a geometrical basis and their positions not refined. Updating of the solvent structure was performed automatically using ARP (Lamzin & Wilson, 1993) in an iterative manner. The water molecules most likely to be wrong were identified and removed and new sites added.

Step	Package	B factors	Hydrogens	Double conf.	R free
1	PROLSQ + ARP (solvent)	Iso	No	No	Yes
2	- " -	Iso	Yes	No	Yes
3	- " -	Iso	Yes	Yes	Yes
4	SHELXL + ARP (solvent)	Iso	Yes	Yes	Yes
5	- " -	Aniso	Yes	Yes	Yes
6	- " -	Aniso	Yes	Yes	All data
7	SHELXL full / block matrix	Aniso	Yes	-	All data

Figure 4. Refinement scheme employed for a protein at atomic resolution.

	R factor 95 % of data	R free 5 % of data	ΔR factor	ΔR free
Initial model	31.1	31.1		
Most of solvent	17.7	20.9	13.4	10.1
Hydrogens	16.9	20.0	0.8	0.9
Double conf.	16.7	19.8	0.2	0.2
Anisotropic B's	11.7	15.0	4.0	4.8
More double conf.	10.9	14.3	0.8	0.7

Figure 5. Refinement of native RNase at 1.2 Å using 95 % of the data.

Automatic determination of the optimum difference electron density threshold and real space fit were employed. All solvent sites were modelled as fully occupied.

Refinement comprised several steps, Figures 4, 5. Introduction of most solvent sites provided the most substantial reduction in R factor, to about 13 %. Placing hydrogens reduced it by a further 1 %. Building of residues in double conformations had little effect with isotropic B factors but gave about a 1 % reduction with anisotropic. Refinement of anisotropic B factors provided 4 %. Reduction in Rfree was essentially the same as for R. Refinement of the complex with 2′-GMP followed the same protocol as that for the native.

Anisotropic refinement allowed location of about 20 % more solvent sites, building of more side chains in double conformations and correcting residue 72. According to the amino acid sequence it was cysteine and in previously refined structures of RNase Sa was modelled in two conformations. In the present refinement the CB-SG distances in both conformers showed significant discrepancy from the target value of 1.808 Å. The difference electron density showed a peak between the SG and CB atoms and a hole at the SG position, clearly indicating that the actual bond length is shorter. Furthermore there are two well defined water molecules at hydrogen bond distances. Cys72 was replaced by Thr72 which fits nicely to the electron density and the distances for its side chain fit well to target values. This clarified long but unsuccessful attempts to use cysteine specific mercury compounds for preparation of isomorphous derivatives.

	Native 1.20 Å	Complex 1.15 Å
R free (%), 5 % of data	14.3	15.5
R factor (%), 95 % of data	10.9	11.2
More cycles with all data		
R factor (%), all data	10.6	10.9

Figure 6. Last steps of refinement using all data.

	Matrix inversion	SigmaA
Isotropic (Rfactor 16.7 %)		
All atoms		0.150
Anisotropic (Rfactor 10.9 %)		
Main chain	0.028	
Side chain	0.043	
Protein total	0.036	
Waters	0.093	
All atoms	0.051	0.050

Figure 7. Coordinate error estimate (Å) for native RNase refined at 1.2 Å with isotropic and anisotropic temperature factors.

All data were used in the last cycles of the refinement, Figure 4. One might expect that if extra data are included and the observations to parameters ratio increases, the conventional R factor should also increase. In practice it is the other way round: the R factor falls by 0.3 % for both structures, Figure 6. Thus use of incomplete data (even if only a small percentage of reflections is randomly omitted) limits the convergence of the refinement even at atomic resolution. This clearly shows one problem of using Rfree as a cross-validation parameter: omission of part of the data (to check whether overfitting has occurred) actually results in a degree of overfitting.

INSINUATIONS

Inversion of the block matrices obtained for anisotropically refined models and the σ_A plot (Read, 1986) were used to estimate the coordinate error, Figure 7. The r.m.s. error for main-chain atoms of both molecules in the two structures was about 0.02 Å. The overall r.m.s. coordinate error for protein atoms was 0.03 Å. The highest inaccuracies were around the most "problematic" parts of the structure, e.g. the Ser3 carbonyl oxygen, which is highly disordered. Both structures with anisotropic atomic temperature factors are essentially 3 times more accurate than structures with isotropic factors. This reflects the quality and demonstrates the advantage of atomic resolution X-ray data.

For most of the structure data to atomic resolution allow correct identification of the atomic type directly from the density map. The electron density interpolated at the atomic centres for each atom type is approximately proportional to the atomic number. The sulphurs of two sulphate anions bound to the protein follow the same dependence even for a temperature factor as high as 80 Å^2.

The two independent molecules in both models are similar with an r.m.s. deviation between CA atoms of 0.38 Å. The maximum displacement of 1.5 Å corresponds to the surface loop Ala62 to Thr64. All differences between molecules A and B may be ascribed to different crystal contacts. Differences between native and complex structures are caused by the presence of inhibitor. After formation of the complex the surrounding atoms of the active site move so that the active site cleft is more closed while the rest of the structure remains unchanged.

Isotropic refinement of the structure was carried out using PROLSQ. 385 waters were identified. Refinement was repeated using SHELXL-93 with the same starting model but with a solvent continuum. Only 288 waters were found, Figure 8. 248 pairs of waters lay within 0.5 Å of one another and 271 within 1.1 Å. Distances between corresponding water sites correlate highly with their temperature factors. High B value waters may not be important for understanding the function of the protein but they do contribute to the quality of the model. The extra waters in the model refined with PROLSQ mostly have high temperature factors and probably mimic the solvent continuum. When more cycles of SHELXL-93 refinement were carried out with the solvent continuum option turned off, the number of solvent sites substantially increased. Thus two different programs with different minimisation function (F's or F^2's) gave essentially the same result. Comparison of models refined isotropically and anisotropically shows that positions of solvent sites systematically differ by at least 0.1 Å.

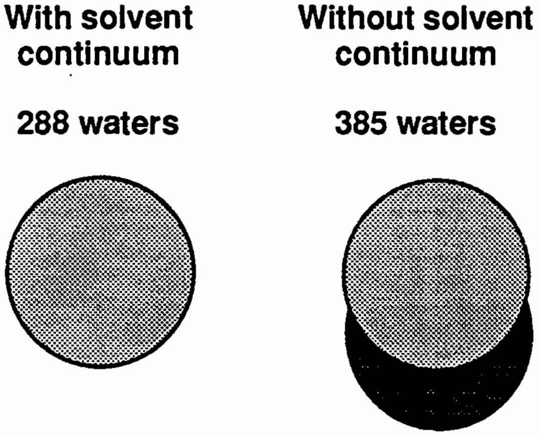


Figure 8. Isotropic refinement of RNase native at 1.2 Å with and without solvent continuum.

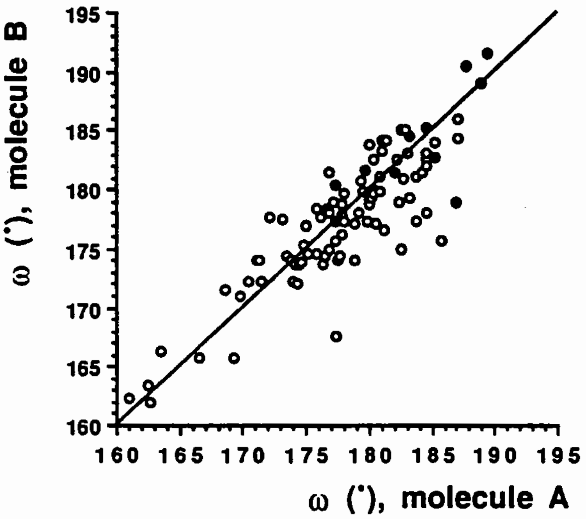


Figure 9. Peptide ω angles for native RNase. Glycines and prolines are shown as filled circles.

During refinement only weak restraints on peptide planarity were used. Peptides appeared to be not quite planar: the ω angle had a mean value of 178.0°, with a standard deviation of

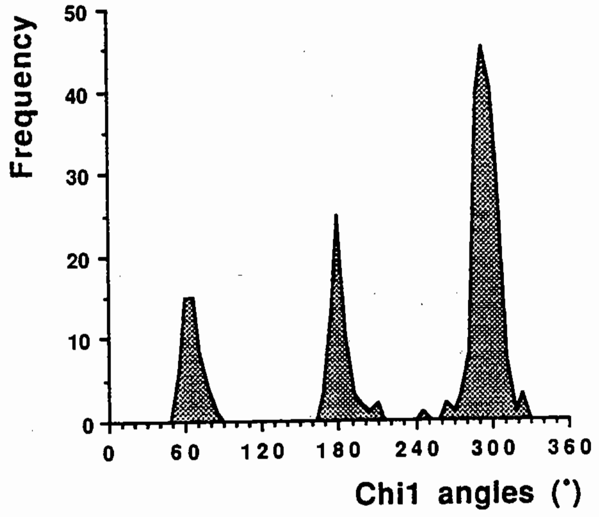
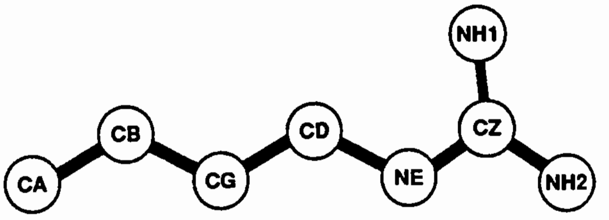


Figure 10. Staggered χ_1 angles.



	4 proteins at atomic resolution 28 contributors	Engh & Huber
CB-CG mean (Å)	1.485	1.520
CB-CG sigma (Å)	0.040	0.030
CG-CD mean (Å)	1.526	1.520
CG-CD sigma (Å)	0.031	0.030

Figure 11. Arginine side chain.

5.8°. Some peptides deviate from planarity by up to 20°. The correlation between ω angles in the A and B molecules fit well to a straight line, Figure 9. The r.m.s. deviation between ω angles in the A and B molecules is 3.0°. Proline and glycine residues are shown separately and are well clustered with average values of 181° and 183° respectively, higher than 180°. These "special" residues are not responsible for the shift of the average angle (178°) for all residues. The average ω angles for all residues excluding glycines and prolines is 177°.

A plot of staggered χ_1 angles (not restrained) shows three peaks at 64° (7), 180° (5) and -64° (9) corresponding to the three possible rotamers, Figure 10. The mean values and standard deviations (in brackets) were derived by fitting three Gaussian functions to the distribution. The rotamer preferences generally agree with those derived from proteins at lower resolution (McGregor *et al.*, 1987) but the mean values are slightly different and standard deviations about 3 times smaller. These could be related to the relatively small sample size (304 contributors for the two RNase structures) but almost certainly reflect the high quality of the models refined at atomic resolution.

Weak restraints were applied for bonded and angle distances. Nevertheless, at atomic resolution the X-ray contribution was dominant and distances were expected to be different from the library. For example, the arginine side chain stereochemistry differs from the target set derived from small molecules, Figure 11. CB, CG and CD atoms were treated as equivalent with sp^3 hybridisation. Two RNase Sa models and a few other proteins refined at atomic resolution gave 28 contributors well defined in the electron density. The standard deviations derived from this limited sample are comparable to the small molecule set. CB-CG mean distance is apparently shorter as derived from proteins. The t-test probability that CB-CG and CG-CD mean distances derived from proteins are different is 0.99986. More structures are needed (and will be provided) for an improved library for protein stereochemistry.

CONCLUSION

Atomic resolution data have been collected for more than 20 proteins at EMBL Hamburg alone. RNase Sa is one of these. The goal of such studies is to bridge the gap between small and large molecule crystallography and to explore the detailed stereochemistry of proteins. Small- and macromolecular crystallography will converge in the future. We are the best (Scaramanga, 007).

REFERENCES

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F.Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535-542.

Blundell, T.L. & Johnson, L.N. (1976). In *Protein crystallography.* Acad. Press, NY, pp. 248-249.

Brünger, A. T. (1993). Assessment of phase accuracy by cross validation: the free R value. Methods and application. *Acta Crystallogr.* **D49**, 24-36.

CCP4 (1994). Collaborative Computational Project Number 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr.* **D50**, 760-763.

Engh, R.A. & Huber, R. (1991). Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallogr.* **A47**, 392-400.

Konnert, J.H. & Hendrickson, W.A. (1980). A restrained-parameter thermal-factor refinement procedure. *Acta Crystallogr.* **A36**, 344-350.

Lamzin, V.S. & Wilson, K.S. (1993). Automated refinement of protein models. *Acta Crystallogr.* **D49**, 129-147.

Leslie, A.G.W. (1992). Recent changes to the MOSFLM package for processing film and image plate data. *CCP4 and ESF-EACMB newsletter on protein crystallography* **26**.

McGregor, M.J., Islam, S.A. & Sternberg, J.E. (1987). Analysis of the relationship between side-chain conformation and secondary structure in globular proteins. *J. Mol. Biol.* **198**, 295-310.

Otwinowski, Z. (1993). DENZO: an oscillation data processing program for macromolecular crystallography. *Yale University, New Haven, USA.*

Read, R.J. (1986). Improved Fourier coefficients for maps using phases from partial structures with errors. *Acta Crystallogr.* **A42**, 140-149.

Sevcik, J., Zegers, I., Wyns, L., Dauter, Z. & Wilson, K.S. (1993). Complex of ribonuclease Sa with cyclic nucleotide and a proposed model for the reaction intermediate. *Eur. J. Biochem.* **216**, 301-305.

Sheldrick, G.M. (1993). SHELXL-93, program for crystal structure refinement, *University of Göttingen,* Germany.

Shlyapnikov, S.U., Both, V., Kulikov, V.A., Dementiev, A.A., Sevcik, J. & Zelinka, J. (1986) Amino acid sequence determination of guanyl-specific ribonuclease Sa from *Streptomyces aureofaciens. FEBS Lett.* **209**, 335-339.

Acknowledgements

This work was supported in part by the EU Biotechnology grant BIO2-CT-92-05-24.

VALIDATING PROTEIN STRUCTURES:
FROM CONSISTENCY CHECKING TO QUALITY ASSESSMENT

S. J. Wodak, Joan Pontius, Alexei Vaguine and Jean Richelle
Université Libre de Bruxelles
Unité de Conformation de Macromolecules Biologiques
CP160/16, P2, Ave. P. Héger, B1050 Bruxelles BELGIUM

1. INTRODUCTION

Progress in genetic engineering, X-ray crystallography, Nuclear Magnetic Resonance spectroscopy and the advent of cheap and powerful computers, has brought about an exponential growth of data on protein 3D structures. Managing the information on protein sequence and structure has become one of the major challenges of modern molecular biology. This is a particularly complex problem in the case of protein 3D structures, because the latter contain a wealth of complex information, some of which must be derived from the atomic coordinates. Meeting this formidable challenge requires efficient ways of storing, cross referencing and accessing the structural information, which can be collectively referred to as 'databases'. Such databases will only be useful if the information and data they contain are consistent and as error free as possible, hence the crucial role of methods and procedures for validating the information and the data. This applies in particular to the atomic coordinates of macromolecules. Owing to the lack of atomic resolution in X-ray and NMR experiments, the data they provide may not be sufficient to define a model accurately enough. The molecular models thus always represent a compromise between the fit to the experimental data and to our knowledge of chemistry. Procedures and criteria for assessing the quality of these models, both overall and in specific portions of the model, are hence of prime importance. In this note, we describe some of the concepts that underlie these procedures and summarize work performed by the teams in the European BIOTECH project dedicated to this subject.

2. PROCEDURES FOR ASSESSING THE QUALITY OF ATOMIC MODELS

Structure validation refers to the procedures for assessing the quality of atomic models of macromolecules. These include proteins, nucleic acids and non polymer groups (heterogens), with the latter encompassing small organic and inorganic molecules, modified amino-acids and nucleic acids, as well as sugar polymers. Three types of procedures make up the ideal validation process. One comprises procedures that can be performed on isolated structures by comparison with basic information on chemical structure and naming conventions stored in protein sequence and small molecule databases, and 'specialized' dictionaries, also called the External Reference Files (ERF's). The second involves comparisons against standard values derived from surveys of the best available macromolecular structures (f.e. those with highest resolution and most extensively refined). The third comprises procedures which involve assessing the quality of the experimental data – such as the structure factors modules (for X-ray structures) or the distance constraints (for NMR structures) – and the agreement of the atomic coordinates with these data.

2.1. *Verification of naming conventions, chemical structure, and covalent geometry*

This is done by checking compliance with information stored in specialized dictionaries, the ERF's. The information checked at this stage includes the IUPAC naming conventions for the standard protein and nucleic acid monomers and for non-standard groups, the chemical structures of the monomers and heterogens, and their atomic connectivities. For proteins, prior to this, the amino-acid sequence needs to be checked against information stored in protein sequence databases such as SWISS-PROT and PIR. In cases of new heterogens for which information is not available in the ERF's, checking their chemistry and nomenclature requires searching for relevant molecular templates in small molecule databases such as the Cambridge Small molecule Databank (CSD), against which they can then be compared.

The next step involves verifying the covalent geometry. This comprises checking the bond length and angles of polymer and non-polymer groups as well as the chirality and planarity of certain sub–groups. These checks perform comparisons against standard values derived from small molecules for proteins [Engh & Huber, 1991], from surveys of the Nucleic acid DataBase (NDB), for the nucleotide groups [Parkinson et al., in preparation] or from the CSD for other heterogens. These standard values are all stored in ERF's.

2.2. *Comparisons against standard values derived from surveys of other macromolecular structures*

This involves comparing various geometric and energetic parameters of the analyzed structure with standard values derived from the best available structures of other macromolecules (highest resolution, most extensively refined; with highest number of distance constraints, etc.). Among the most commonly surveyed parameters are the backbone and side chain torsion angles, and the distortions from planarity and tetrahedral geometry. Other interesting parameters can also be used to assess 3D models, and validation criteria based on their analysis are in development in several laboratories (see below). These include the number of 'bad' contacts per residue, H–bond energies and geometries, atomic/ residue volumes; volumes of empty cavities; atomic/residue accessibilities, residue environment following Bowie et al. [1991], non–bonded energies (or non–bonded contacts) between polymer atoms as well as with ligands and solvent molecules.

2.3. *Assessing the agreement of the model with the experimental data*

Given the lack of atomic resolution in macromolecular structure determination, be it by X–ray diffraction or NMR spectroscopy, the derived 3D model is always a compromise between the experimental data and what we know to be a reasonable chemical structure.

Clearly, therefore, validation procedures which rely solely on compliance with standard values of geometric or energetic parameters derived from known macromolecular structures or small molecules, only address part of the quality assessment problem. Furthermore, they invariably introduce bias into the validation procedure due to the requirement to comply with best existing models or structural information, irrespective of the quality and availability of experimental data. Quality measures must therefore include criteria based on the agreement with the experimental data. For crystal structures such criteria may involve computing Omit Maps, or evaluating Free R factors. For structures determined by NMR this would involve analyzing agreement with distance and/or dihedral angle constraints.

Furthermore, ideally, structure validation procedures should *combine* criteria based on agreement with experimental data with those based on compliance with standard geometric or energetic parameters, derived from databases.

3. EUROPEAN BIOTECH (FRAMEWORK III) PROJECT ON: 'INTEGRATED PROCEDURES FOR RECORDING AND VALIDATING RESULTS OF 3D STRUCTURE STUDIES OF BIOLOGICAL MACROMOLECULES'

In the following we list the partners of the project and highlight achievements made so far.

3.1. *The Project partners are:*

- EMBL (Hamburg–Germany): Keith Wilson (Project Coordinator) and Victor Lamzin.
- University of York (UK): Eleanor Dodson and Garib Murshudov
- University College London (UK): Janet Thornton and Roman Laskowski
- Bijvoet Center, Utrecht (The Netherlands): Robert Kaptein and Ton Rullman
- EMBL (Heidelberg, Germany): Chris Sander, Gert Vriend and Robert Hooft
- Université Libre de Bruxelles (Belgium): Shoshana Wodak, Joan Pontius, Jean Richelle and Alexei Vaguine
- University of Uppsala: Alwyn Jones and Gerard Kleywegt

3.2. *Current achievements*

3.2.1. *Verification of naming conventions, chemical structure and covalent geometry*

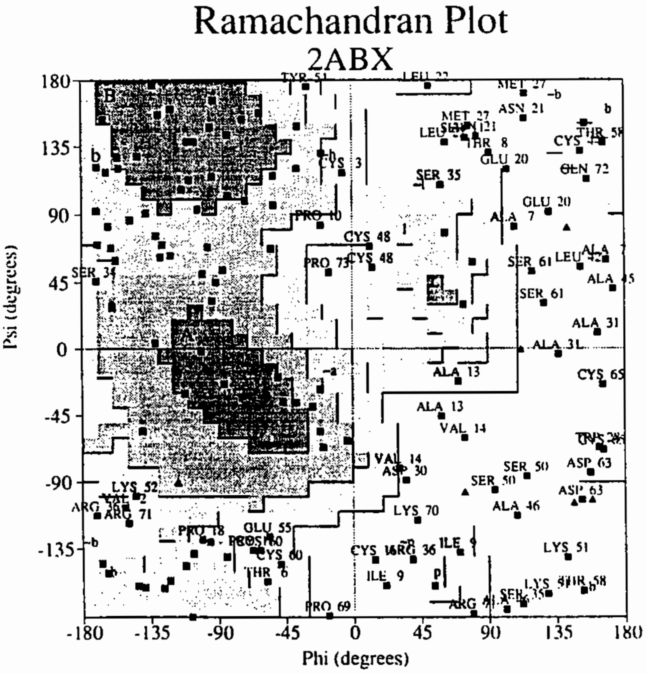
Here developed procedures include those available in the packages of WHAT–IF (EMBL, Heidelberg), BRUGEL/SESAM (ULB, Brussels), IDITIS and PROCHECK

(UCL, London). WHAT–IF [Vriend, 1990] and BRUGEL [Delhaise et al., 1985] are general purpose molecular modelling packages which apply many of the validation checks as part of the procedures of entering new atomic coordinates into the package for further analysis. SESAM [Huysmans et al., 1991] and IDITIS [Thornton and Gardner, 1989] are relational databases of macromolecular 3D structures. They include validation procedures as part of their data entry stream. PROCHECK [Morris et al., 1992: Laskowski et al., 1993] is a stand–alone validation package, which has become increasingly popular; it validates essentially only geometric parameters (see below).

3.2.2. Comparisons against standard values derived from surveys of other macromolecular structures

Procedures based on this type of validation are available in PROCHECK, WHAT–IF, and SURVOL (ULB, Brussels). The latter is a stand–alone software package which evaluates atomic volumes and compares them to standard values derived from surveys of other macromolecules [Alard, 1991; Pontius et al., in preparation] (see below). New procedures and validation criteria are being developed for NMR structures by the UCL–London and Utrecht groups. Criteria to analyze and validate H–bonding interactions with solvent are being developed at the EMBL–Heidelberg and UCL–London, and methods for analyzing and assessing non–bonded interactions are being designed at UCL–London, the EMBL–Heidelberg and ULB–Brussels.

Figures 1–3 illustrate structure validation output provided by PROCHECK for the PDB entry of 2ABX (α–bungrotoxin crystal structure), whose atomic model seems to have some problems, and which we therefore denote as 'outlier': Figure 1 maps the ϕ,ψ values of this entry onto the Ramachandran map in which regions are shaded according to their likelihood to be populated in good protein structures of the database. We see that many of the ϕ,ψ values correspond to disallowed regions, an indication that the quality of this model is poor. Figure 2 (a,b) summarizes in a pictorial fashion the detected geometry distortions in 2ABX relative to the standard Engh & Huber [1991] values. Figure 3 summarizes several of the quality assessment measures used in PROCHECK, for the same structure. It confirms that the quality of this structure is well below that expected for good structures at the same resolution (2.3Å).



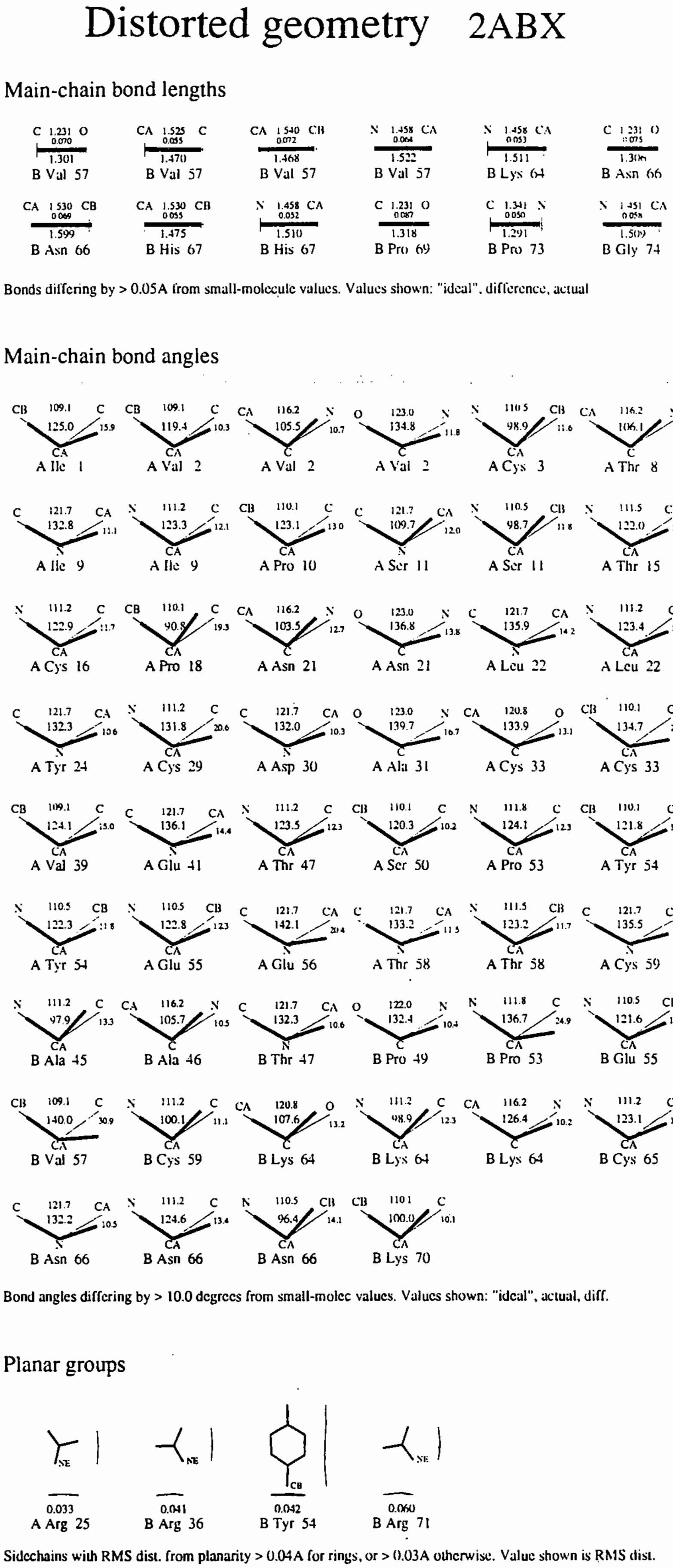
Plot statistics

Residues in most favoured regions [A,B,L]	18	14.8%
Residues in additional allowed regions [a,b,l,p]	45	36.9%
Residues in generously allowed regions [~a,~b,~l,~p]	29	23.8%
Residues in disallowed regions	30	24.6%
	----	------
Number of non-glycine and non-proline residues	122	100.0%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown as triangles)	8	
Number of proline residues	16	

Total number of residues	148	

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.

Figure 1: ϕ, ψ, values of 2ABX (α–bungrotoxin crystal structure) onto the Ramachandran map



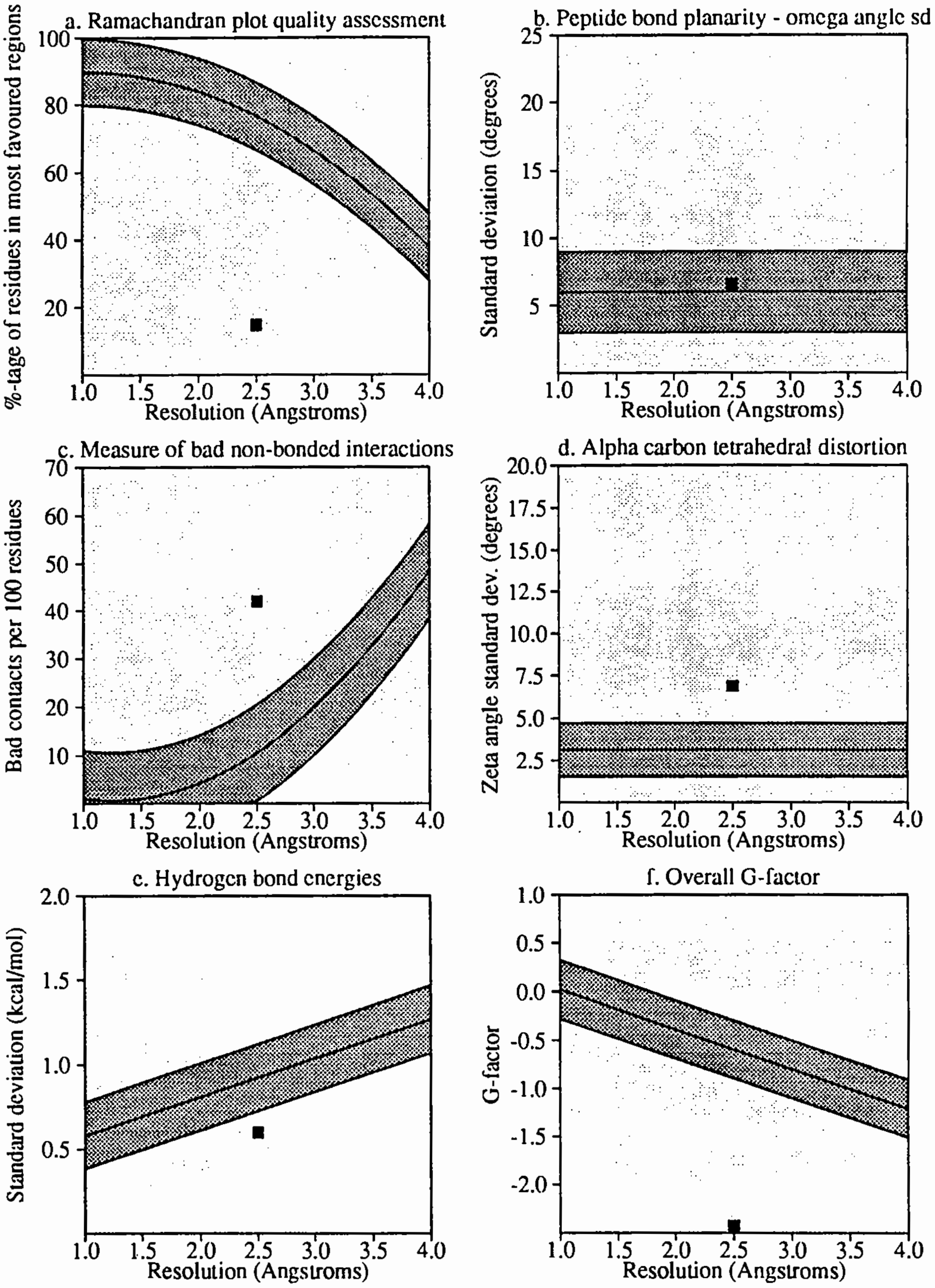
Bonds differing by > 0.05A from small-molecule values. Values shown: "ideal", difference, actual

Bond angles differing by > 10.0 degrees from small-molec values. Values shown: "ideal", actual, diff.

Sidechains with RMS dist. from planarity > 0.04A for rings, or > 0.03A otherwise. Value shown is RMS dist.

Figure 2: Geometry distortions of 2ABX (α−bungrotoxin crystal structure) relative to standard Engh & Huber [1991] values. (a) main−chain bond lengths & main−chain bond angles. (b) planar groups

Main-chain parameters 2ABX



Plot statistics

Stereochemical parameter	No. of data pts	Parameter value	Comparison values Typical value	 Band width	No. of band widths from mean	
a. %-tage residues in A, B, L	122	14.8	76.6	10.0	-6.2	WORSE
b. Omega angle st dev	146	6.5	6.0	3.0	0.2	Inside
c. Bad contacts / 100 residues	62	41.9	10.5	10.0	3.1	WORSE
d. Zeta angle st dev	140	6.9	3.1	1.6	2.3	WORSE
e. H-bond energy st dev	36	0.6	0.9	0.2	-1.6	BETTER
f. Overall G-factor	148	-2.4	-0.6	0.3	-6.1	WORSE

Figure 3: Quality assessment measures used in PROCHECK, for the entry 2ABX (α−bungrotoxin crystal structure)

Figures 4–6 illustrate the work done in Brussels on structure validation using as criteria atomic volumes, and implemented in the program SURVOL. This work involved computing the atomic volumes in a subset of highly resolved and refined protein structures in the database, and analyzing the distributions of these volumes for different atomic types defined according to their chemical nature and bonded environment. This required among other things choosing a consistent set of atom types and deriving the atomic radius for each type, based on the analysis of pair–correlation functions in structures of the database.

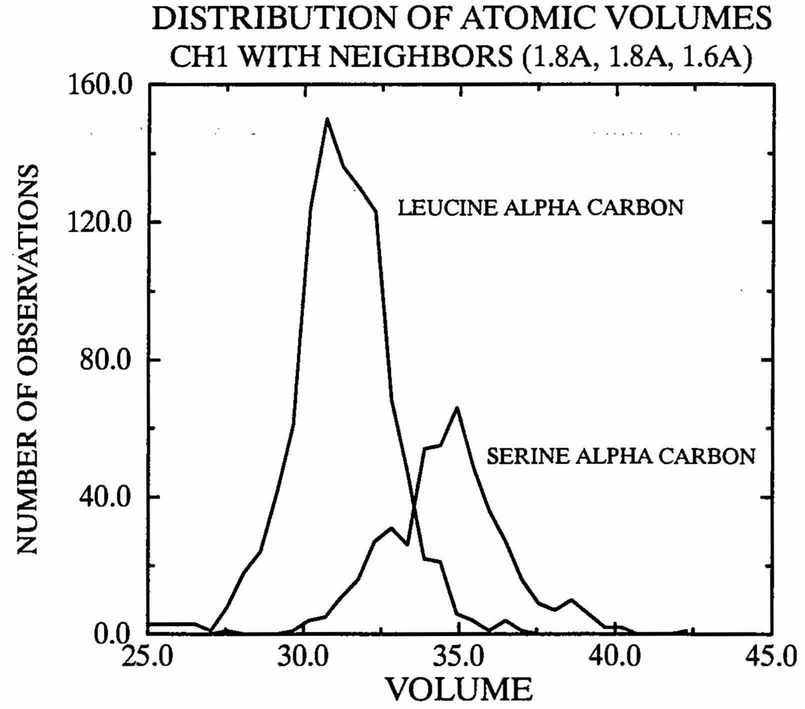


Figure 4: Volume distributions for the Cα atom in Leu and Ser residues

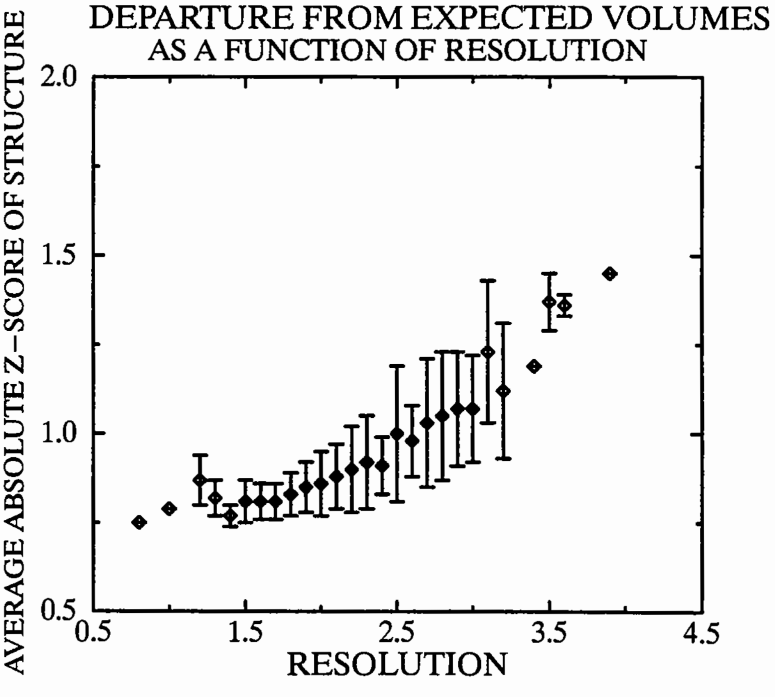


Figure 5: Average global Z–score and its standard deviation, as a function of resolution

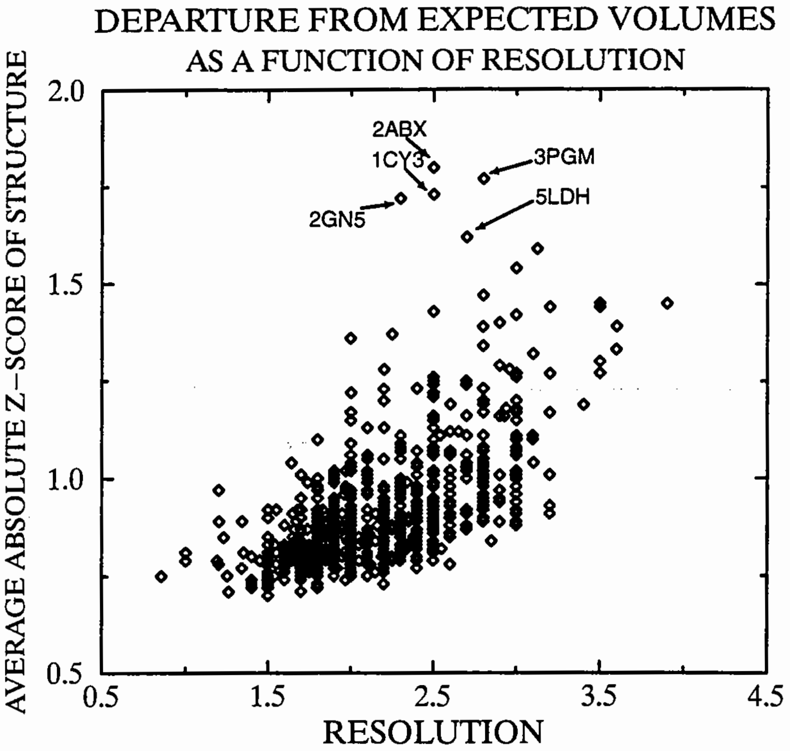


Figure 6: Individual global Z−scores of the analyzed proteins as a function of resolution

Figure 4 displays the volume distributions for the Cα atom in Leu and Ser residues respectively. It illustrates that though the chemical nature of the considered atom (Cα), and the radii of the atoms bonded to it (1.8Å, 1.8Å, 1.6Å) are the same in the 2 residues, the corresponding volume distributions are very different, indicating an influence from groups two bonds away (a OH in Ser, and a CH2 in Leu). Different standard volumes were therefore assigned to the Cα in these residues. Similar considerations were used in deriving target values and standard deviations for other atom types. Validation based on the evaluation of atomic volume then consists in computing the volume of atoms in a given structure and comparing these volumes to those expected from the computed distributions. For each atom a Z−score is computed. This Z−score is defined as the difference between its volume and the expected target volume, divided by the standard deviation of the volume distribution of the appropriate atom type. Atoms which have Z−scores above a certain threshold value, may correspond to problem regions in the protein model. The model as a whole can be evaluated based on the average absolute Z−score of its atoms. This global Z−score has been computed for proteins determined at different resolutions ranging from 1 to 3.7Å. Figure 5 displays the average global Z−score and its standard deviation, as a function of resolution. It shows that this Z−score increases as the resolution becomes poorer, and hence as the model quality deteriorates. Figure 6 displays the individual global Z−scores of the analyzed proteins as a function of resolution. 'Problem' structures such as 2ABX or 2GN5, have global Z−scores much higher than average and thus appear as outliers on this figure.

3.2.3. Validation of isolated molecules against X−ray data

Work done in this area addresses the compromise aspect of the atomic model of a protein, mentioned above. It concerns the development of simple procedures for validating models derived by X−ray diffraction based on compliance with standard geometry, as well as with experimental data. These procedures involve model optimizations with different target functions aimed at independently maximizing the agreement with experimental data (observed structure factors) or with geometric parameters (bond distances and angles).

Figure 7 illustrates this concept. The 'working model', representing a given X−ray structure which has just come off the mill of a standard refinement package, is subject to 2 independent refinement processes. In one it is refined to maximize agreement with the ideal geometry parameters, without taking into account agreement with the X−ray data, which leads to the 'ideal model'. In the other, the working model is refined to maximize agreement with the X−ray data, without any geometric constraints, which leads to the 'X−ray model'. Comparisons of the atomic coordinates of the ideal and X−ray models is seen to reveal problem regions in the structure: regions where agreement with the ideal geometric parameters and the X−ray data is at odds.

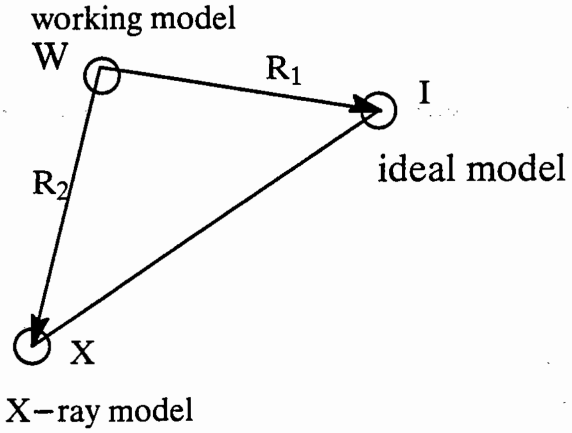


Figure 7: Model optimization with different target functions. R_1 denotes a refinement which maximizes agreement with the geometric constraints, completely neglecting agreement with X−ray data; R_2 denotes a refinement which maximizes agreement with the X−ray data not taking into account any geometric constraints. The definitions of W, X and I are given in the text.

Figure 8 illustrates preliminary results obtained in experiments which apply this procedure to RNase Ap1, a 1.7Å resolution structure solved by Poliakov and colleagues [personal communication]. 3 independent experiments were conducted. In experiment 1, a 'good' working model of RNase Ap1 (W), refined at 2.5Å resolution with isotropic B−factors, is subjected to 10 cycles of least squares refinement, only with geometric constraints, leading to model I, and only with X−ray data, leading to model X, respectively. The rmsd of the atomic coordinated between these models is 0.20Å. In experiment 2, a 'correct' model of the same protein, obtained from the 'good' model by random displacements of individual atoms, undergoes the same treatment, leading to models whose rmsd is now 0.25Å. Finally in experiment 3, a 'wrong' model, derived from the 'good' model by rotating the entire structure in the unit cell, is subjected to the double refinement, yielding 2 models whose coordinates display a significantly larger rmsd of 0.53Å

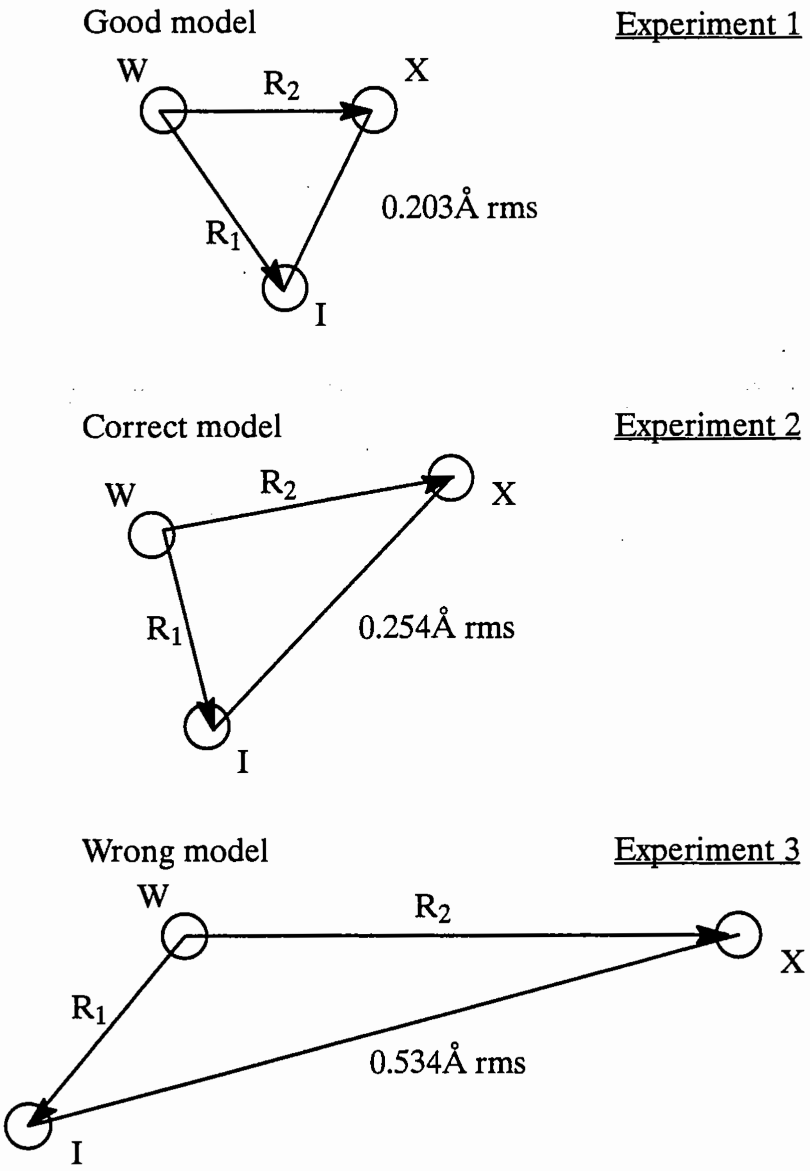


Figure 8: Refinement experiments done on RNase Ap1 as a test case. Experiments 1, 2 and 3 are described in the text. The displayed rms deviations indicate differences in atomic coordinates between the resulting modes.

Figure 9 illustrates how the rmsd between models I and X, in experiment 2, varies along the sequence (Fig. 9a), and that this variation correlates well with that of the B–factors (Fig. 9b).

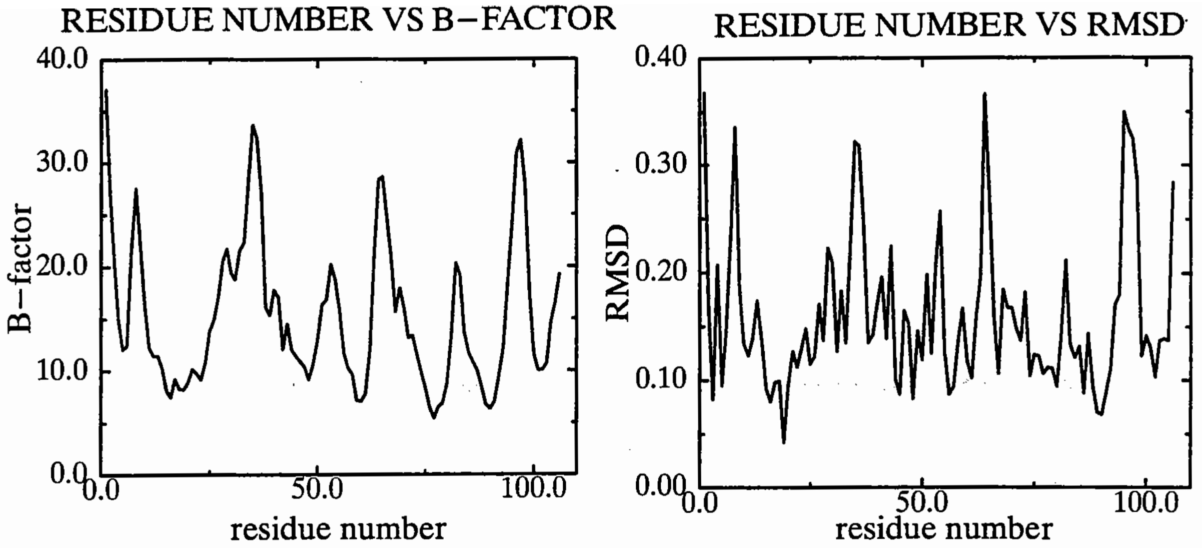


Figure 9: rms deviations along the sequence between models I and X obtained in experiment 2 (b) in comparison with the B−factor along the sequence (a)

These preliminary results indicate that the double refinement approach is a useful tool for quantifying the combined agreement with X−ray data and geometry constraints, not only for the structure as a whole, but also in specific regions in the polypeptide. Work is in progress to apply this approach to a larger set of examples.

4. CONCLUDING REMARKS

The developments described above are only the beginning of an important field of activity that will surely expand in coming years.

Future developments will concern:

- New file standards, for exchanging and representing macromolecular data. File standards such as the CIF (Crystallographic Interchange File) developed under the auspices of the IUCr, and others, will contain more complete information on the macromolecular structure and experimental data; it will also have a better representation of this information, which is machine readable and parsable, making consistency checking and error proof reading much easier.
- The development of procedures and links to other databases, such as sequence databases, and small molecule databases.
- The design of better quality measures based on comparison with experimental X−ray and NMR data respectively.
- Quality assessment methods for disordered models and for multiple sets of conformations such as those generated by NMR studies.
- Ways of annotating quality assessment results in archived structures should also be developed in full cooperation with the scientific community which generates the structural information.

These are only a few examples of the directions in which activities in this field will go, and should not be considered as an exhaustive list.

5. NETWORK ACCESS TO EXISTING VALIDATION PROCEDURES DEVELOPED BY THE BIOTECH VALIDATION PROJECT

Several of the model validation procedures described above can be accessed through the network at the following addresses:

- http://www.sander.embl−heidelberg/check/index.html (PROCHECK;WHAT−IF)
- http://www.ucmb.ulb.ac.be/~joan/survol.html (SURVOL)

REFERENCES

Alard, P., PhD thesis, Université Libre de Bruxelles (1991)

Bowie, J.U., Lüthy, R. and Eisenberg, D. A method to identify protein sequences that fold into a known three–dimensional structure. Science, *253* (1991) 164–170

Delhaise, P., Van Belle, D., Bardiaux, M., Alard, P., Hamers, P., Van Cutsem, E. and Wodak, S. Analysis of data from computer simulations on macromolecules using the CERAM package. J. Mol. Graph., *3* (1985) 116–119

Engh, R..A. and Huber, R. Accurate bond and angle parameters for X–ray protein structure refinement. Acta Cryst., *A47* (1991) 392–400

Huysmans, M., Richelle, J. and Wodak, S.J. SESAM: a relational database for structures of macromolecules. Proteins, *11* (1991) 59–76

Laskowski, R.A., Moss, D.S. and Thornton, J.M. Mainchain bond lengths and bond angles in protein structures. J. Mol. Biol., *231* (1993) 1049–1067

Morris, A.L., MacArthur, M.W., Hutchinson, E.G. and Thornton, J.M. Stereochemical quality of protein structure coordinates. Proteins, *12* (1992) 345–364

Parkinson, G., Vojtechovsky, J., Clowney, L. and Berman, H.M. New parameters for refinement of nucleic acid containing structures. (1995) (in preparation)

> The new topology and parameter files for X–plor can be accessed via the NDB W3 server: http://ndbserver.rutgers.edu:80, and can be found under the keyword *refinement*

Pontius, J.U., Wodak, S.J. and Richelle, J. Quality assessment of protein 3D structures using standard atomic volumes. (1995) (in preparation)

Thornton, J.M. and Gardner, S.P. Protein motifs and data–base searching. TIBS, *14* (1989) 300–304.

Vriend, G. WHAT IF: a molecular modelling and drug design program. J. Mol. Graph., *8* (1990) 52–56

Protein Structures in Solution Viewed by NMR.

Lorna J. Smith and Christopher M. Dobson.

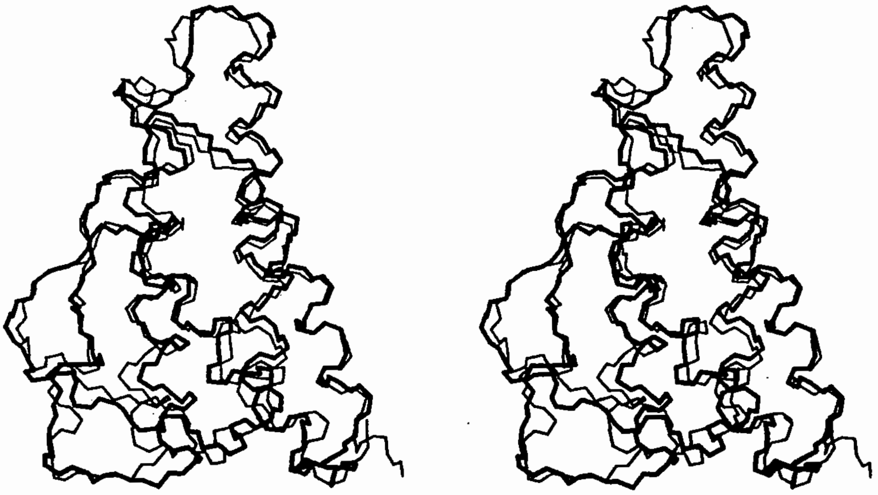
Oxford Centre for Molecular Sciences and New Chemistry Laboratory,
University of Oxford, South Parks Road, Oxford OX1 3QT ENGLAND.

Recent developments in nuclear magnetic resonance (NMR) techniques, particularly involving the use of isotopically labelled samples and multidimensional heteronuclear experiments, have meant that NMR has emerged as a powerful method for determining the structures of proteins, even those with a molecular weight in excess of 20 kDa, in solution (Wagner, 1990; Clore & Gronenborn, 1991a). In cases where the structure of the same protein has been solved independently by NMR and crystallographic techniques there has been considerable interest in comparisons of these structures (Billeter et al., 1989; Clore & Gronenborn, 1991b; Berndt et al., 1992; Billeter, 1992; Smith et al., 1994). Such comparisons can provide insight into the level of similarity between protein structures in the solution and crystalline states, as well as providing a method for identifying any errors that may arise from either technique. In general, comparisons have indicated the close similarity of the overall structures of proteins in solution and in crystals, particularly in the protein core, although in some cases apparently significant and real differences have been observed. Examples of these include aspects of the quaternary structure of interleukin-8 (Baldwin et al., 1991; Clore & Gronenborn, 1991c) and the metal coordination site of metallothionein (Furey et al., 1986; Schultze et al., 1988). However, the NMR studies have also revealed a wealth of information about more subtle differences between proteins in crystals and solution, and particularly about the nature and significance of protein dynamics (Karplus & McCammon, 1981; Dobson, 1993). Here, we illustrate this with some of our results from NMR studies of human interleukin-4 and hen lysozyme, including comparisons with crystallographic data for these proteins.

Comparison of structures determined by NMR and X-ray diffraction techniques.

For the helical cytokine human interleukin-4 (IL-4) the opportunity for a particularly interesting comparison arose (Smith et al., 1994) as four independent structures determined in different laboratories, two by NMR spectroscopy (Smith et al., 1992; Powers et al., 1992) and two by X-ray diffraction (Wlodawer et al., 1992; Walter et al., 1992), were published within a few months of each other; an overlay of two of these structures is shown in Figure 1. All four structures have an identical overall fold, a left-handed four helix bundle with an up-up-down-down connectivity, although regions with differing conformation were found. These were concentrated for the main chain in the long loops that run the length of the bundle and in the N- and C-termini. In an NMR structure determination an ensemble of structures is calculated, each member of which satisfies the NMR data and the covalent geometry requirements of the protein. When the ensemble of NMR structures for IL-4 was analysed, the atoms in these same loop and terminal regions of the main chain were found to have large rmsd's when the different structures were compared with each other. In the case of the side chains, greater differences were found between the four independent structures than had been observed for the main chain, and greater rmsd values were found within the ensemble of NMR structures (Redfield et al., 1994a), particularly for those side chains on the protein surface.

Figure 1: Stereodiagram showing a backbone superposition of two structures of human interleukin-4, one determined using NMR techniques (thin line) (Smith et al., 1992) and the other by X-ray diffraction (bold line) (Wlodawer et al., 1992).



The observation of these differences prompts the question of whether the disorder observed, for example for the loops and surface side chains, in the ensemble of NMR structures is merely a result of insufficient or ambiguous structural data for the regions concerned or whether it indicates that these regions are mobile. NMR techniques provide the potential to answer this question as they give the opportunity to probe the dynamics of a protein in a residue specific manner through the observation of averaging of NMR parameters and exchange effects, and through the analysis of relaxation measurements. In the following section we provide some specific examples of this and compare the results with crystallographic data.

Probing protein dynamics by NMR techniques.

a) Averaging of NMR parameters.

If there is mobility in the protein structure, either involving librations around a single conformation or interconversion between populations of distinct multiple conformations, then the observed values of NMR parameters such as chemical shifts and coupling constants will be a population weighted mean of the values expected for the individual conformers, as long as the rate of interconversion between the conformers is fast on the NMR time scale. Fast exchange for homonuclear coupling constants corresponds, for example, to a rate of interconversion between the contributing conformers that is greater than approximately 20 Hz. If such averaged NMR parameters can be recognised they can provide direct evidence for motional processes occurring in the protein in solution. Recognition and interpretation of motional averaged NMR parameters is simplest in the case of spin-spin coupling constants, the most commonly measured values being $^3J_{HN\alpha}$ and $^3J_{\alpha\beta}$. These can be related to the main chain ϕ and side chain χ_1 angles respectively via the Karplus relationship (Karplus, 1959), although because of the degeneracy of this relationship measurement of a single coupling constant value does not provide a unique value for the torsion angle. Recent developments in heteronuclear NMR techniques have, however, been increasing the range of protein coupling constants that can be measured (Wagner, 1990) giving the potential for almost all the torsion angles in proteins to be probed.

The structure of the enzyme hen lysozyme has been determined in solution by NMR techniques (Smith et al., 1993). Detailed comparisons with crystal structures of the protein have shown the close similarity of the structure of this protein in solution and in crystals. For lysozyme we have measured both $^3J_{HN\alpha}$ (for 106 residues) and $^3J_{\alpha\beta}$ (for 57 residues) coupling constants (Smith et al., 1991). These values have been compared with those predicted from crystal structure ϕ and χ_1 angles using the Karplus equation. For $^3J_{HN\alpha}$ there is found to be a close agreement between the experimental and calculated values (Figure 2a). This shows the close similarity between the backbone conformations of the protein in solution and in crystals, and suggests that any large scale motions of the backbone in solution are very limited (Smith et al., 1991).

For $^3J_{\alpha\beta}$ (Figure 2b) it is immediately obvious that the agreement between the experimental and predicted values is less good. For residues where the side chain adopts a single staggered conformation one would expect, for residues with two β protons, two small $^3J_{\alpha\beta}$ values or one large and one small value. For many residues this is indeed observed (see examples in Table IA) and there is good agreement with the values predicted from the crystallographic χ_1 angles, although for residues with a small coupling constant the experimental value is generally larger than the calculated one, while for residues with a large coupling constant the experimental value is usually smaller than the calculated one. These differences reflect small fluctuations about the average torsion angle. For residues where multiple χ_1 conformations are adopted in solution, the observed $^3J_{\alpha\beta}$ values will be averaged and if this is extensive both values will lie in the range 6-8 Hz. This motional averaging of the $^3J_{\alpha\beta}$ values is observed for 16 of the 57 residues of lysozyme studied (see examples in Table IB). All of these residues, with one exception Val 99, are on the protein surface and, as the crystal χ_1 angles for these residues are usually those for specific staggered rotamers, there is a poor agreement between the calculated and experimental coupling constant values.

Figure 2: Comparison of the experimental coupling constants for hen lysozyme with those calculated from the angles in the tetragonal type 2 crystal structure (Handoll, 1985) of the protein using the Karplus relationship. In (a) the main chain $^3J_{HN\alpha}$ coupling constants (excluding those for glycine residues) are shown and in (b) the side chain $^3J_{\alpha\beta}$ coupling constants (Smith et al., 1991). For residues where the β-methylene protons have not been stereospecifically assigned the calculated $^3J_{\alpha\beta}$ coupling constants are plotted against the experimental values so as to give the best agreement.

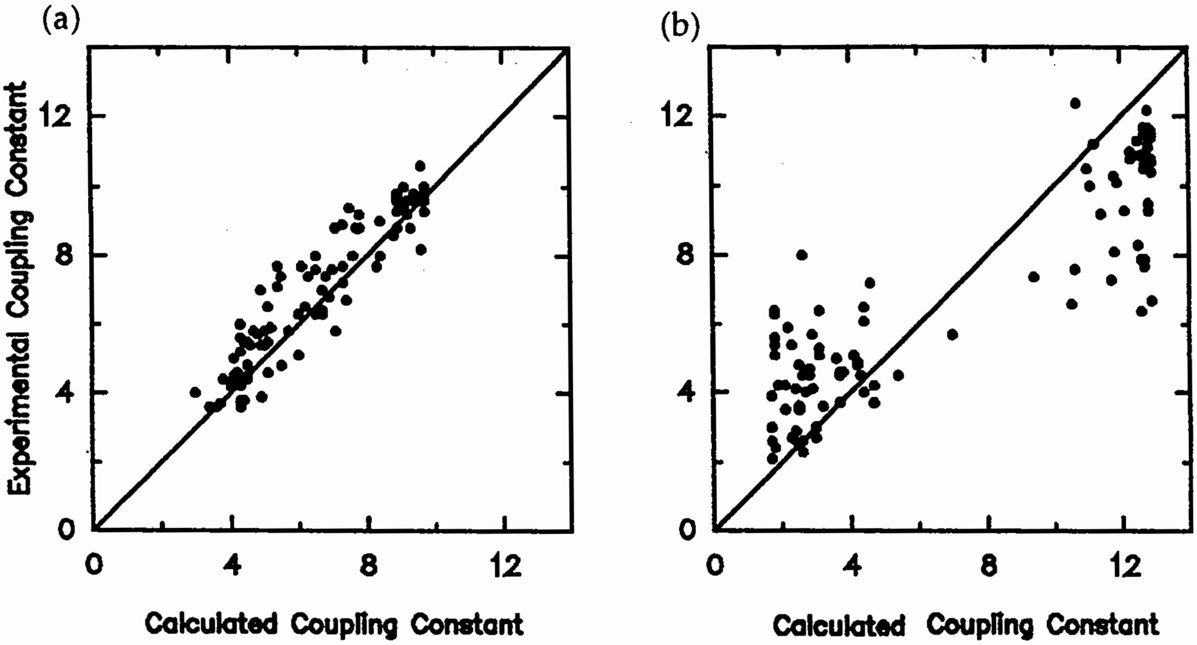


Table I. Examples of $^3J_{\alpha\beta}$ coupling constants and corresponding χ_1 angles in the tetragonal type 2 and triclinic crystal structures of hen lysozyme (data from Smith et al., 1991).

A) Residues that occupy a single staggered χ_1 conformation in solution.

Residue	$^3J_{\alpha\beta}$ (Hz)‡	χ_1 tetragonal	χ_1 triclinic
Cys 6	3.5, 11.5	-68.2°	-67.5°
Tyr 20	2.3, 11.7	-172.8°	179.7°
Asn 39	10.8, 4.5	-172.3°	-172.5°
Cys 64	2.7, 4.6	63.6°	67.3°
Trp 123	2.9, 10.6	-69.0°	-68.9°

B) Residues that occupy multiple χ_1 conformations in solution

Residue	$^3J_{\alpha\beta}$ (Hz) ‡	χ_1 tetragonal	χ_1 triclinic
Glu 7	6.7, 6.4	-177.6°	-65.7°
Arg 45	6.9, 6.7	-175.1°	-66.4°
Val 99	6.3	80.3°	171.1°
Asp 101	6.6, 5.6	-151.4°	-97.9°
Arg 125	7.9, 6.1	-52.9°	166.4°

‡ In each case the coupling constant for the β proton whose resonance has the high chemical shift is listed followed by that for the β proton with the lower chemical shift.

It is interesting to see whether there is any evidence for such side chain mobility in crystal structures of the protein. For hen lysozyme, 50% of the residues with averaged $^3J_{\alpha\beta}$ coupling constants in solution have γ atom temperature factors in the crystal that are greater than 30Å^2 (compared with 7% of residues that adopt single χ_1 conformations). Although none of the crystal structures reported for hen lysozyme resolves multiple side chain conformations, comparison of the χ_1 conformations of structures of different crystal forms reveal different structures for some side chains. For example, comparison of tetragonal and triclinic structures of lysozyme shows that 9 of the 16 residues with motionally averaged $^3J_{\alpha\beta}$ values are found to adopt significantly different conformations in the two crystal forms (see examples in Table IB). Hence, side chain mobility in solution can be reflected in disorder or different conformations in different crystal structures although the mobility in solution is more extensive than analysis of a single crystal structure would suggest (Smith et al., 1991).

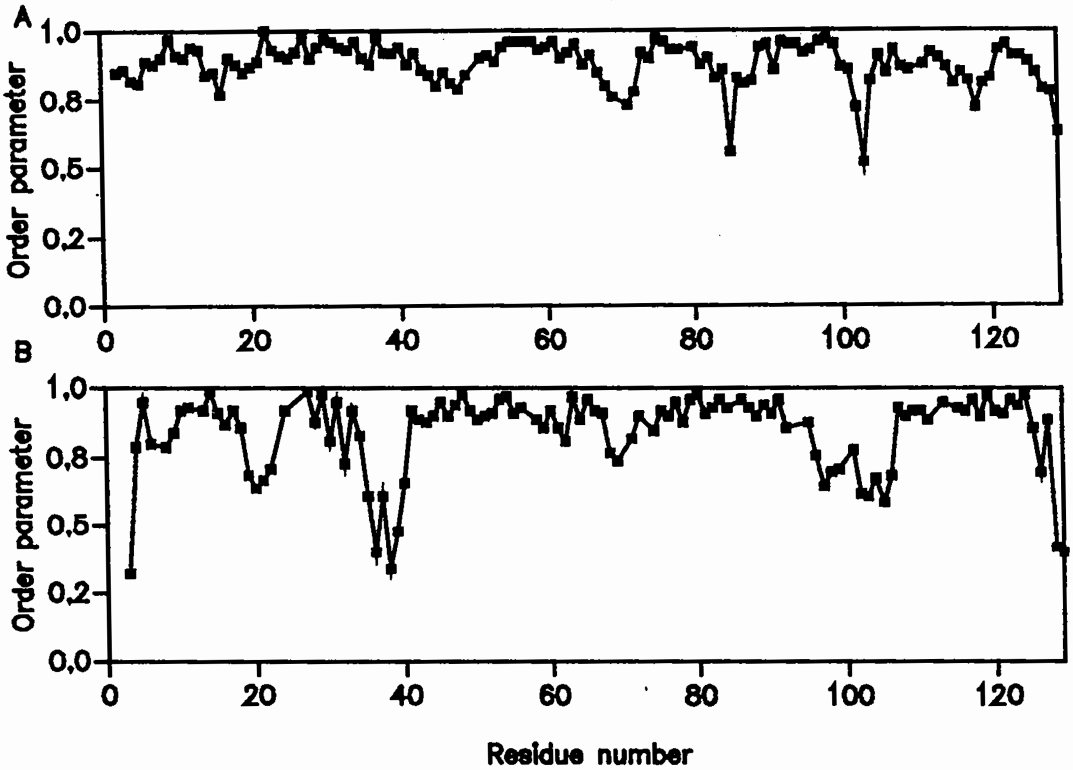
$^3J_{\alpha\beta}$ coupling constants measurements have not yet been made for IL-4. However, the result for hen lysozyme that 55% of the residues studied with a solvent accessibility greater than 60% adopt multiple χ_1 conformations in solution is closely similar to those found for other proteins where similar studies have been performed including FK506 binding protein (Xu et al., 1992) and ribonuclease A (Rico et al., 1993). This suggests that mobile surface side chains are a common feature of proteins in solution and implies that the disorder of surface side chains seen in the ensemble of NMR structures of IL-4 and the differences in side chain conformation between the independent IL-4 structures does not merely arise from lack of structural data but reflects the dynamics of the side chains.

b) Relaxation measurements.

The study of the relaxation behaviour of uniformly ^{15}N labelled proteins provides a powerful technique for studying main chain dynamics through probing the motion of the ^{15}N-^1H vectors (Wagner, 1993). In general ^{15}N T_1 and T_2 relaxation rates and ^1H-^{15}N NOE values are measured and the data analysed using the model free approach of Lipari and Szabo (1982), order parameters being extracted for the individual amide groups in the sequence. These order parameters give a measure of the degree of spatial restriction of the ^1H-^{15}N bond vector, very restricted fluctuations on a fast picosecond timescale (i.e. involving motions faster than the rotational correlation time of the molecule) giving a value close to 1.0.

The ^1H-^{15}N order parameters resulting from these ^{15}N relaxation experiments performed for two proteins, hen lysozyme (Buck et al., 1994) and human IL-4 (Redfield et al., 1992), in our laboratory are shown in Figure 3. For hen lysozyme only two residues, Ser 85 and Asn 103, have order parameters significantly less than 0.7; there is therefore no evidence for mobile loop regions or hinge bending motions in this protein. For IL-4, in contrast, although high order parameters (average 0.88) are seen throughout the helices, lower order parameters (S^2 0.3-0.8) are observed for the terminal regions of the sequence and the long AB and CD loops, except in the region where they are linked to each other by a short region of antiparallel β-sheet structure. These regions with low order parameters are amongst the regions where differences have been observed within the ensemble of NMR structures and between the four independent IL-4 structures indicating that in these cases the disorder and different conformations reflect the presence of internal motions.

Figure 3: Main chain ^1H-^{15}N order parameters (S^2) versus sequence number for hen lysozyme (A) (Buck et al., 1994) and human interleukin-4 (B) (Redfield et al., 1992).

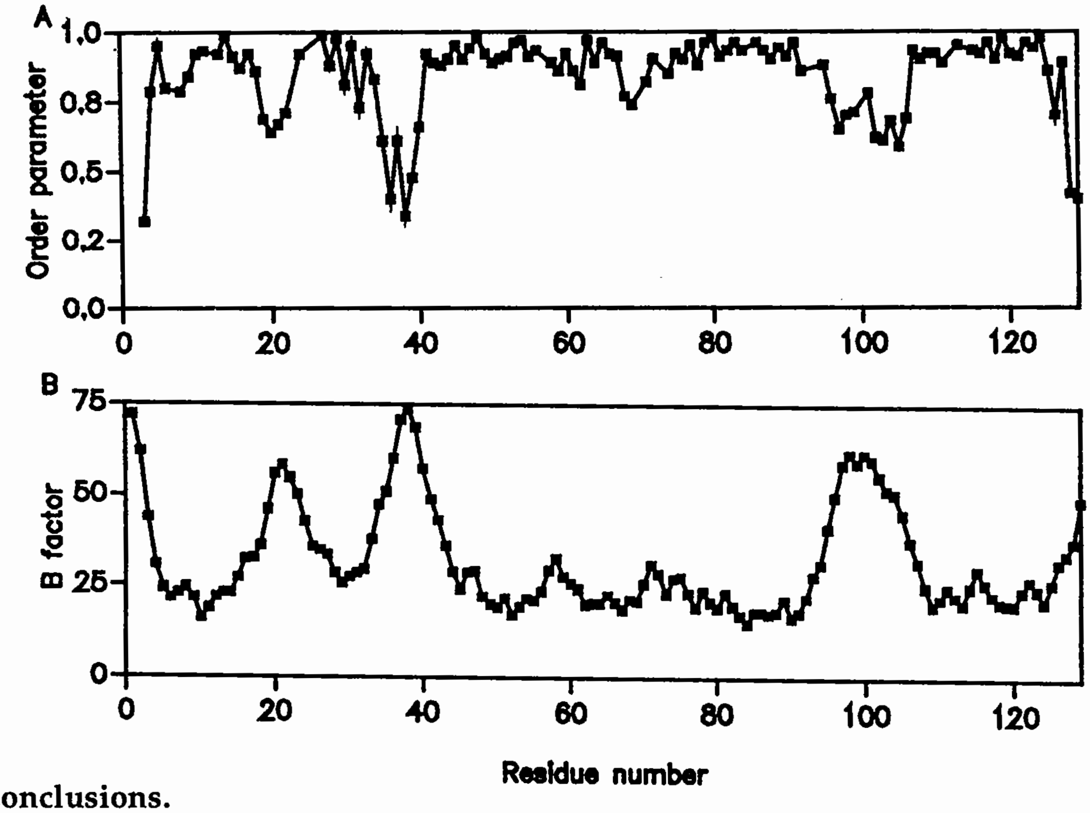


The dynamics of side chains as well as the main chain can be probed by ^{15}N (Trp indole NH, Asn and Gln NH_2, Arg NεH) relaxation studies. In the case of hen lysozyme the results of these experiments show that there is a wide variety of motional behaviour for different side chains, the experimental order parameters ranging from 0.2-0.9 (Buck et al., 1994). However, in agreement with the results of the $^3J\alpha\beta$ coupling constant analysis, a correlation between the side chain order parameters and solvent accessibility has been identified, a residue such as Asn 27 with a relatively buried side chain (20% solvent accessibility) having a high order parameter (0.79) while an exposed side chain such as that of Asn 77 (93% accessibility) having a much lower value of 0.22.

There has been considerable interest in attempts to assess the relationship between ^{15}N order parameters and crystallographic B factors (Powers et al., 1993). In some cases, such as human interleukin-4 (see Figure 4), there is a high degree of correlation, regions with low order parameters in solution exhibiting high N atom temperature factors in crystal structures of the protein

(Smith et al., 1994). In other proteins, however, there are residues where there is little or no correlation between these two parameters, reflecting the fact that, for example, lattice contacts can restrict the dynamics of regions in crystals that are mobile in solution and crystallographic B factors can include contributions from disorder resulting from mobility on a slower timescale than that reflected in order parameters calculated in the conventional way. The possibilities for comparisons of this type are becoming further enhanced now that side chain as well as main chain ^{15}N order parameters are being measured.

Figure 4: Comparison of crystallographic temperature factors and solution order parameters for human interleukin-4. A) and B) show the ^{1}H-^{15}N order parameters for human IL-4 (Redfield et al., 1992) and the temperature factors of the main chain nitrogen atoms from the crystal structure of Wlodawer et al. (1992) respectively.



Conclusions.

As a result of the development of NMR techniques for the determination of protein structures we have been able to gain considerable insight into the behaviour of folded proteins in solution. In general the structures adopted in solution have been found to be very similar to those determined for protein crystal but an increased disorder is found on the protein surface in solution. Indeed close packing of the polypeptide chain seems to be required for mobility to be restricted and a single conformation adopted. This is in accord the results of structural studies of linker regions connecting domains in larger proteins, peptides and partly folded and unfolded proteins where the extent of close packing is considerably reduced or absent and far more extensive mobility and conformational disorder is observed (Perham, 1991; Dobson, 1992; Redfield et al., 1994b).

The presence of mobility and conformational disorder on the surface of proteins is of considerable importance as binding and recognition sites all involve surface residues. Here it appears that, rather than being a property that decreases their functional effectiveness, in many cases biology may have actually exploited the dynamic nature of proteins (Dobson, 1993). For example, the mobility of a binding site may enable the regions concerned to carry out a dynamic search for the required site on the binding partner so speeding up recognition and increasing adaptability. In addition, on binding the flexibility then helps the complementarity of the surfaces to be maximised (induced fit) so improving the strength of binding. The mobility of binding sites has another potentially important thermodynamic consequence. Generally on binding much of the mobility is lost leading to an unfavourable entropy change. In some cases this is thought to be important for reducing the free energy of molecular association so providing reversibility of binding without loss of specificity (Searle et al., 1992).

In this article we have demonstrated the considerable insight that can be gained from combining NMR data with that from crystallographic techniques. In some cases such comparisons have been taken even further and joint refinements have been performed using both NMR and X-ray data (Shaanan et al., 1992). There has also been considerable interest in whether NMR data can be productively combined with data from other experimental or theoretical techniques. In this regard comparison of NMR data with results from theoretical MD simulations appears to be particularly promising and has the potential to increase considerably our understanding of the dynamic properties of proteins (Dobson & Karplus, 1986; Chandrasekhar et al., 1992; Eriksson et al., 1993).

Acknowledgements.

This is a contribution from the Oxford Centre for Molecular Sciences which is supported by the U.K. Engineering and Physical Sciences Research Council, the Biotechnology and Biological Sciences Research Council and the Medical Research Council. The research of C.M.D. is supported in part by an International Research Scholars award from the Howard Hughes Medical Institute. L.J.S. is a Royal Society University Research Fellow.

References.

Baldwin, E.T., Weber, I.T., St. Charles, R., Xuan, J-C., Appella, E., Yamada, M., Matsushima, K., Edwards, B.F.P., Clore, G.M., Gronenborn, A.M. & Wlodawer, A. (1991) Proc. Natl. Acad. Sci. USA., *88*, 502-506.

Berndt, K. D., Güntert, P., Orbons, L. P. M., & Wüthrich, K. (1992) J. Mol. Biol. *227*, 757-775.

Billeter, M. (1992) Quaterly Reviews of Biophysics *25*, 325-377.

Billeter, M., Kline, A. D., Braun, W., Huber, R., & Wüthrich, K. (1989) J. Mol. Biol. *206*, 677-687.

Buck, M., Redfield, C., Boyd, J.,MacKenzie, D. A., Jeenes, D. J.,Archer, D. B., & Dobson, C. M. (1994) Biochemistry, *in press.*

Chandrasekhar, I., Clore, G. M., Szabo, A., Gronenborn, A. M., & Brooks, B. R. (1992) J. Mol. Biol. *226*, 239-250.

Clore, G. M. & Gronenborn, A. M. (1991a) Ann. Rev. Biophys. Biophys. Chem. *20*, 29-62.

Clore, G. M. & Gronenborn, A. M. (1991b) J. Mol. Biol. *221*, 47-53.

Clore, G. M. & Gronenborn, A. M. (1991c) J. Mol. Biol. *217*, 611-620.

Dobson, C. M. (1992) Cur. Opin. Struc. Biol. *2*, 6-12.

Dobson, C. M. (1993) Current Biology *3*, 530-532.

Dobson, C. M., & Karplus, M. (1986) Methods in Enzymol. *131*, 362-389.

Eriksson, M. A. L., Berglund, H., Härd, T., & Nilsson, L. (1993) Proteins *17*, 375-390.

Furey, W. F., Robbins, A. H., Clancy, L. L., Winge, D. R., Wang, B. C., & Stout, C. D. (1986) Science *231*, 704-710.

Handoll, H. H. G. (1985) D. Phil. thesis, University of Oxford.

Karplus, M. J. Chem. Phys. *30*, 11-15 (1959).

Karplus, M., & McCammon, J. A. (1981) CRC Crit. Rev. Biochem. *9*, 293-349.

Lipari, G., & Szabo, A. (1982) J. Amer. Chem. Soc. *104*, 4546-4559.

Perham, R. N. (1991) Biochemistry *30*, 8501-8512.

Powers, R., Garrett, D.S., March, C.J., Frieden, E.A., Gronenborn, A.M., & Clore, G.M. (1992) Science *256*, 1673-1677.

Powers, R., Clore, G. M., Garrett, D. S., & Gronenborn, A. M. (1993) Journal of Magnetic Resonance. Series B *101*, 325-327.

Redfield, C., Boyd, J., Smith, L. J., Smith, R. A. G., & Dobson, C. M. (1992) Biochemistry *31*, 10431-10437.

Redfield, C., Smith, L. J., Boyd, J., Lawrence, G. M. P., Edwards, R. G., Gershater, C. J., Smith, R. A. G., & Dobson, C. M. (1994a) J. Mol. Biol. *238*, 23-41.

Redfield, C., Smith, R. A. G., & Dobson, C. M. (1994a) Nature Structural Biology *1* 23-29.

Rico, M., Santoro, J., Gonzalez, C., Bruix, M., Neira, J. L., & Nieto, J. L. (1993) Applied Magnetic Resonance *4*, 385-415.

Shaanan, B., Gronenborn, A.M., Cohen, G.H., Gilliland, G.L., Veerapandian, B., Davies, D.R., & Clore, G.M. (1992). Science, *257*, 961-964.

Schultze, P., Worgotter, E., Braun, W., Wagner, G., Basak, M., Kagi, J. H. R., & Wüthrich, K. (1988) J. Mol. Biol. *203*, 251-268.

Searle, M. S., Williams, D. H., & Gerhard, U., J. Amer. Chem. Soc. (1992) *114*, 10697-10704.

Smith, L. J., Sutcliffe, M. J., Redfield, C., & Dobson, C. M. (1991) Biochemistry *30*, 986-996.

Smith, L.J., Redfield, C., Boyd, J., Lawrence, G.M.P., Edwards, R.G., Smith, R. A. G., & Dobson, C.M. (1992) *J.* Mol. Biol. *224*, 900-904.

Smith, L. J., Sutcliffe, M. J., Redfield, C., & Dobson, C. M. J. Mol Biol. (1993) *229* 930-944.

Smith, L. J., Redfield, C., Smith, R. A. G., Dobson, C. M., Clore, G. M., Gronenborn, A. M., Walter, M. R., Nagabushan, T. L., & Wlodawer, A. (1994) Nature Structural Biology *1*, 301-310.

Walter, M. R., Cook, W. J., Zhao, B. G., Cameron, R. P., Ealick, S. E., Walter, R. L., Reichert, P., Nagabushan, T. L., Trotta, P. P. & Bugg, C. E. (1992) J. Biol. Chem.. *267*, 20371-20376.

Wagner, G. (1990) Progress in NMR Spectroscopy *22*, 101-139.

Wagner, G. (1993) Cur. Opin. Struc. Biol. *3*, 748-754.

Wlodawer, A., Pavlovsky, A. and Gustchina, A. (1992) FEBS Lett., *309*, 59-64.

Xu, R. X., Olejniczak, E. T., & Fesik, S. W. (1992) FEBS Lett. *305*, 137-143.

MAKING THE MOST OF YOUR MODEL
CCP4 Study Weekend 6-7 January 1995

Comparative Structure Analysis

Liisa Holm and Chris Sander
Protein Design Group
European Molecular Biology Laboratory
D-69012 Heidelberg, Germany

Summary and Introduction

For physicists, proteins are linear polymers which fold into complicated three-dimensional shapes. In spite of much recent progress, theoretical understanding of folding principles is still incomplete. Most of our knowledge of protein structures is based on experimental structure determination. It is usual to discuss and describe a newly solved protein structure in comparative terms: What are the similarities to the corpus of previously known structures and which features are unique ? Here, we give an overview of three computational methods for comparative structure analysis. **(1) Model quality:** empirically observed regularities in protein structures can be useful in assessing the quality (reasonableness) of a model as it emerges from the electron density map. Atomic solvation preference profiles are a computationally inexpensive diagnostic tool for checking sequence placement relative to chain tracing (program SolPref).
(2) Substructures: secondary structure, folding motifs and structural domains are abstractions developed during the era of visual inspection of molecular models. The use of quantitative physical criteria allows more objective definition of, e.g., domain borders (program Puu). **(3) Search for structural similarities:** an important development of recent years is the arrival of a new generation of automated computer algorithms that allow routine comparison of a protein structure with the rapidly growing database of all known structures (our contribution: program Dali). Such structure database searches are already used daily and they are beginning to rival sequence database searches as a tool for discovering remote evolutionary connections. Such links between proteins or protein families can, in turn, provide a key to the difficult task of interpreting the structure of a protein in terms of biological function.

Evaluating models using statistical preferences

Faced with the lack of an accurate theory of protein folding, a key aspect in model building methodology is the development of empirical criteria with sufficient discriminatory power to tell a good model from one of lower quality. At an elementary level, the covalent geometry of amino acids yields relatively narrow distributions for bond lengths and angles and backbone torsion angles or side chain rotamers (e.g., Holm and Sander, 1991, 1992b). Statistical correlations between structural state and residue (or atom) type are commonly used to derive effective potentials of mean force, where the preference for a given state is calculated from the 'observed/expected' ratio (e.g., Ouzounis et al., 1993).

The hydrophobic effect has long been known to be important for protein stability, but is difficult to model accurately in molecular simulations. We used the database of known protein structures to derive a novel set of atomic solvation preference parameters for 87 atom types, i.e., all heavy atoms in amino acid side chains (Holm and Sander, 1992a). The environment of atoms was characterized according to the solvent contact model (Colonna-Cesari and Sander, 1990) which is an excluded volume approximation to protein-solvent interaction. The basic idea is that in the first few solvation shells volume not occupied by other protein atoms is filled by water. In more detail, the volume occupancy around a protein atom is calculated as the sum over all volumes of protein atoms in a shell of 6 Å radius weighted with an envelope function that depends on the radial distance from the atom. The solvation factor represents the solvation state of an atom, on a scale from zero to one. Minimal occupancies (solvation factor equal to one) for each atom type were calculated from extended peptide models. Maximal occupancy values (solvation factor equal to zero) were calibrated from known structures, using only residues with less than 4 % relative solvent accessibility. Atomic volumes were taken from Motoc and Marshall (1985).

The solvation preferences were derived using frequencies of occurrence of side chain atom types collected from a database consisting of 63 non-homologous, high-resolution protein structures. The preference of an atom type to occur at each one of eleven bins of solvation factor value was calculated from the 'observed/expected' ratio. The preferences are made additive by taking the logarithm and the sign is chosen so that low (e.g. negative) values are favorable, in analogy with free energy scales. Straight lines were fitted to the data to derive smoothed preferences described by the slope and intercept at zero solvation factor. To apply the preferences to assess a model, the solvation state of each atom is computed and the associated solvation preference values are summed up.

The ability of solvation preference to discriminate between correct and incorrect 3D structures for a given sequence or to identify the correct sequence placement in a given structure was tested among models which have been misfolded in various ways. Backbone coordinates were taken from experimentally known structures or hypothetical models and side chain conformations (in rotamer space) were optimized by an efficient Monte Carlo algorithm using simulated annealing and simple potential functions (Holm and Sander, 1992b). Discrimination by solvation preference was very clear between deliberately misfolded and correct globular models as well as between native-like and non-native-like topologies of combinatorially generated myoglobin models. Due to its statistical nature, the evaluation works best on entire protein models while the identification of incorrect parts of models is more difficult. In favorable cases, locally incorrect chain tracing in a crystal structure may be identified. Figure 1 illustrates the application of the method to two independently determined (different) structures of the DNA-binding gene V protein.

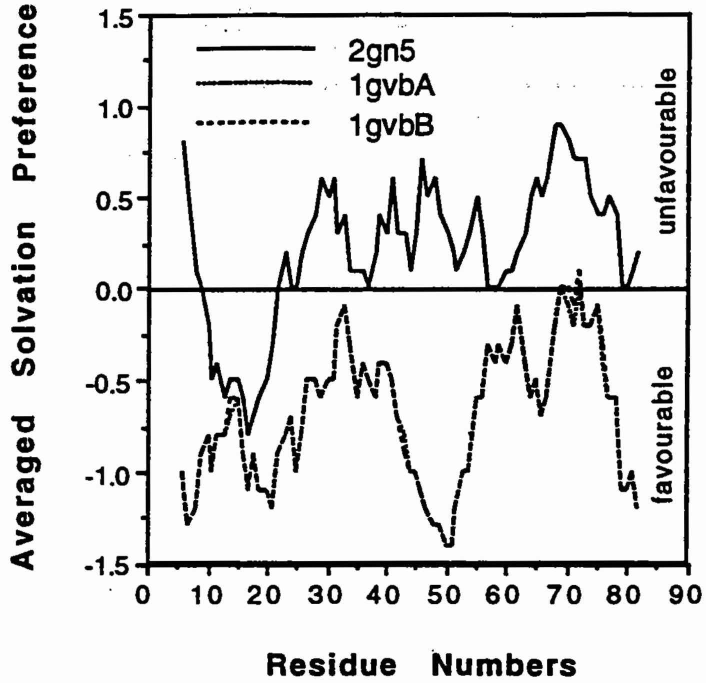


Figure 1.

Atomic solvation preference profiles (Holm and Sander, 1992a) of two Protein Data Bank data sets of the gene V protein. 1gvb (a homodimer with chains A and B) is the minimized average NMR structure by Folkers et al. (1994). 2gn5 is the earlier crystallographic model by Brayer and McPherson (1983), now known to be incorrect in part. The profiles show that threading of the new model (1gvb) places the amino acid side chains in statistically considerably more favorable environments than does the older model. The models differ not so much in the tracing of the backbone (2.5 Å root mean square positional deviation for 82 structurally equivalent C^{α} atoms (Holm and Sander, 1993)) as in the placement of the sequence. Sequence placement relative to the chain trace between 2gn5 and 1gvb is identical only in the segments 42-72 and 80-87. Elsewhere there are shifts of 1,2,3 and 7 residues.

The principal limitation of this method is that it provides only a very rough approximation to energetics. Specific polar and electrostatic interactions are ignored in the present implementation. Main chain atoms are not evaluated at all. A limitation of the misfolded test cases is that the side chains can be optimally fit onto the rigid backbone structure only for the sequence from which the backbone structure was taken. Nevertheless, solvation preference parameters appear overall a remarkably powerful discriminator between incorrectly and correctly folded globular protein models. The technical advantage of the solvent contact formulation over surface area calculations is that the degree of solvation of an atom becomes a particularly simple function of interatomic distances, allowing rapid calculation of solvation-related quantities. A similar solvation concept has been implemented in the Gromos package for molecular dynamics simulations (Stouten et al., 1993).

Description of protein architecture

General patterns of protein structure organization have emerged from studies of hundreds of structures elucidated by X-ray crystallography and nuclear magnetic resonance spectroscopy. In all but the smallest proteins, the polypeptide chain forms several compact, globular units, sometimes loosely connected. Such units are commonly called structural domains, although this definition based on visual inspection is intuitive and therefore rather imprecise. The program Puu (parser for protein unfolding units; Holm and Sander, 1994b) implements an automatic algorithm for identification of structural units by objective, quantitative criteria based on atomic interactions.

A protein may unfold in small bits and pieces (loops, ends) or in large units (structural domains). Let us focus on the second alternative and ask: What are the domains or folding units into which a globular protein separates as it unfolds ? Intuitively, folding units are compact and the interactions between them weak. This intuition is made quantitative in a simple model. Unfolding starts with slow coherent relative motion of the units and mutual rearrangement of solvent and local protein structure near the interface between the units. Gradually, solvent enters into the interface. In the final stages, flexible hinges connect independently solvated, spatially separated units. For this process to occur, the relative motion of the units must be sufficiently slow to allow significant structural rearrangement: the slower, the better. Using a simple harmonic approximation, interdomain dynamics is determined by the strength of the interface and the distribution of masses. As the relative motion of the units occurs on the same time scale as solvent motion, within an order of magnitude, the coupling between the two is strong and even small differences in the time scale may significantly affect the probability of unfolding. Therefore, the main criterion for identifying folding units is the interunit fluctuation time, for which a lower limit can be calculated. For proteins of known 3D structure, the model predicts the most likely decomposition into folding units.

The decomposition of a convoluted 3D structure is complicated by the possibility that the chain can cross over several times between units, which would soon lead to a combinatorial explosion as the number of allowed chain traversals increases. However, the generalized problem of finding a binary partition with any number of cut points in the sequence can be solved making a reasonable approximation. If the residues are grouped by solving an eigenvalue problem for the contact matrix (ignoring mass for the moment), the problem is reduced to a one-dimensional search for all reasonable trial bisections in a band diagonalized contact matrix. The bisection which gives the highest inter-domain fluctuation time is remembered, and used. Recursive bisection yields a tree of putative folding units. Simple physical criteria are then used to identify units that could exist by themselves, based on the ideas of weak interactions between the domains to be separated (large fluctuation time) and strong intra-domain cohesion (compact shape) of each resulting domain after separation.

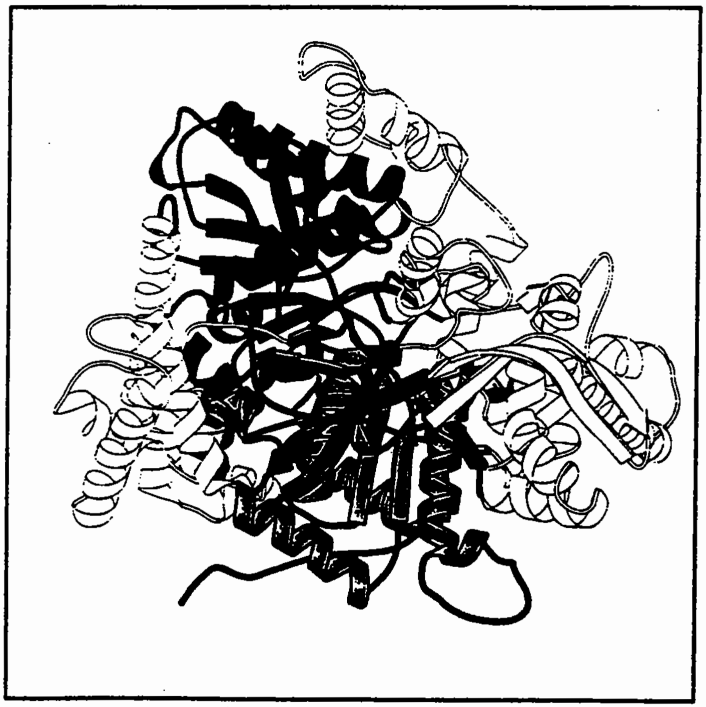


Figure 2.
Decomposition of the 3D structure of glycogen phosphorylase (1gpb, Acharya et al. 1991) into structural domains using the PUU algorithm. The central catalytic core of two α/β domains (dark and light gray ribbon) is externally decorated by a number smaller regulatory domains (white ribbon) formed by excursions of the peptide chain. A nice example of concordance of structural and functional domains.

Tests on a representative set of proteins structures indicated that the units so defined are in good agreement with crystallographers' notion of structural domains (Holm and Sander, 1994b). An example is in Fig 2. As structural domains are basic units of protein structure, function and evolution, the method can be a useful tool in the automatic classification of recurrent folding motifs in newly solved structures.

Structure comparison

Algorithms for structure comparison have to address the full complexity of similarity of shape in 3D space. The notion of structural equivalence becomes increasingly complex with increasing evolutionary distance. The conformation of a point mutant differs from that of the wildtype protein only locally and only by a few tenths of an Ångström. Much larger deviations are observed in pairs of homologous proteins: with increasing sequence dissimilarity, small shifts in the relative orientations of secondary structure elements accumulate and reach several Ångströms and tens of degrees, as described e.g. for the globins. At the largest evolutionary distances, only the topology of the fold or folding motif is conserved, i.e., the relative location of helices and strands and the loop connections between these. Deviations can be even larger and qualitatively different when structural similarity is the result of convergent rather than divergent evolution. In particular, convergent evolution may result in similar 3D folds that differ in the topology of loop connections.

Just as there is much latitude in the basic formulation of the structure comparison problem, many different types of optimization algorithm have been employed. Early computer methods required manual initial alignment or were very slow or limited to close homologues. Increased computer power has paved the way for a new generation of search algorithms that are general, elegant and/or fast (reviewed in Holm and Sander, 1994a).

We have developed a method (implemented in the Dali program, Holm and Sander, 1993) for structure alignment which is based on the exploitation of distance matrices, a rotation and translation invariant representation of three-dimensional structure. The alignment algorithm optimizes a structural similarity score that measures the agreement of all equivalent intramolecular distances in two proteins. More precisely, consider two proteins, labelled *A* and *B*. An alignment consisting of *L* equivalenced residue pairs is evaluated by an additive similarity score of the form

$$(1) \qquad S=\sum_{i=1}^{L}\sum_{j=1}^{L}\phi(i,j) \ ,$$

where i and j label the equivalenced pairs (i^A, i^B) and (j^A, j^B), and ϕ is a measure of the similarity of the C^{α}-C^{α} distances d_{ij}^{A} in protein A and d_{ij}^{B} in protein B. Residues with no equivalent in the other protein do not contribute to the score. ϕ is defined so that smaller distance deviations correspond to higher similarity. Tolerance to spatially extended geometrical distortions is achieved by using relative rather than absolute distance deviations, preventing dominance of long intramolecular distances:

$$(2)\ \phi\ (i,j) = \begin{cases} \left(\theta - \dfrac{|d_{ij}^{A} - d_{ij}^{B}|}{d_{ij}^{*}}\right) w(d_{ij}^{*}),\ i \neq j \\ \theta\ ,\ i = j \end{cases}$$

where d_{ij}^{*} is the arithmetic average of d_{ij}^{A} and d_{ij}^{B}, θ is a constant that determines the threshold of similarity (set to $\theta=0.20$, i.e., 20 % deviation), and the Gaussian damping factor $w(d)=\exp(-d^2/\alpha^2)$ limits the range of the scoring scheme to the radius of a typical domain ($\alpha=20$ Å).

The optimal structural alignment is defined as the set of residue equivalences that maximizes the similarity score. Because of the complex pairwise dependencies, a Monte Carlo procedure is used for optimization. The algorithm is not guaranteed to reach the global optimum, but in practice the method identifies structural similarities in excellent agreement with intuitive notions of structural equivalence. The method allows sequence gaps of any length and free topological connectivity of aligned segments, including matches in reversed chain direction.

Structure searches by human experts or using new algorithms such as Protep, Whatif-Suppos, SSAP, Align, Stamp, Dali for scanning structural databases continue to yield interesting results (reviews: Orengo et al., 1993; Holm and Sander, 1994a). As more and more resemblances and remote evolutionary connections between new and old protein structures are discovered, the growth curve for unique folds rises much less steeply than that for the total number of known structures. Two effects limit the total variety of observed protein folds: physical principles and the evolutionary history of natural proteins. Sequence database searches are a powerful tool in molecular biology because inferences concerning protein structure and function exploit the biological fact that these can be retained over large evolutionary distances. Structure database searches expand the realm of comparative analysis. Experience shows that the chances of finding interesting similarities in the database are already greater than those of a fold being unique. Occasionally biochemical classifications get simplified by the unification of two or more protein families into one superfamily, potentially leading to considerable time savings in the biochemist's laboratory.

Availability of programs and databases

The programs for solvation preference analysis (SolPref, Holm and Sander, 1992a) and domain decomposition (Puu, Holm and Sander, 1994b), as well as other software of interest to molecular biologists, are available by anonymous ftp from ftp.embl-heidelberg.de, in the directory /pub/software. Requests for structure database searches of newly solved crystallographic or solution NMR structures (C^{α} co-ordinates required) by the Dali program may be sent to L. Holm by email (holm@embl-heidelberg.de).

The results of an all-against-all comparison of a sequence-representative set of structures in the Protein Data Bank, i.e., structural alignments and overall classification of protein fold families, are available in the FSSP database (Holm and Sander, 1994c) by anonymous ftp (file transfer protocol) from ftp.embl-heidelberg.de, directory: /pub/databases/protein_extras/fssp. Access to the database is also possible over the World Wide Web (WWW), e.g. using the XMosaic interface; the URL address is http://www.embl-heidelberg.de/databases/protein_extras/fssp. Related databases provided by members of the Protein Design group are available in the same parent directory (sequence families: HSSP; secondary structure definitions: DSSP; domain decomposition: PUU) .

References

Acharya K.R., Stuart D.I., Varvill K.M., Johnson L.N.. Glycogen phosphorylase B: description of the protein structure. World Scientific Publishing Co, Singapore, 1991.

Brayer, G.D. and McPherson, A. J. Mol. Biol., 169 (1983) 565.

Colonna-Cesari, F. and Sander, C. Biophys. J., 57 (1990) 1103-1107.

Folkers, P.J.M, Nilges, M., Folmer, R.H.A, Konings, R.N.H. and Hilbers, C.W. J. Mol. Biol., 236 (1994) 229.

Holm, L. and Sander, C. J. Mol. Biol., 218 (1991) 183-194.

Holm, L. and Sander, C. J. Mol. Biol., 225 (1992a) 193-205.

Holm, L. and Sander, C. Proteins, 14 (1992b) 213-223.

Holm, L. and Sander, C. J. Mol. Biol., 233 (1993) 123-138.

Holm, L. and Sander, C. Proteins, 19 (1994a) 165-173.

Holm, L. and Sander, C. Proteins, 19 (1994b) 256-268.

Holm, L. and Sander, C. Nucl. Acids Res., 22 (1994c) 3600-3609.

Motoc, I. and Marshall, G.R. Chem. Phys. Letters, 116 (1985) 415-419.

Orengo, C.A., Flores T.P, Jones D.T., Taylor W.R., Thornton J.M., Current Biology 3, 131-139 (1993).

Ouzounis, C., Sander, C., Scharf, M. and Schneider, R. J. Mol. Biol., 232 (1993) 805-825.

Stouten, P.F.W., Frömmel, C., Nakamura, H. and Sander, C. Mol. Simul., 10 (1993) 97-120.

Comparison of protein folds and sidechain clusters using algorithms from graph theory.

Peter J. Artymiuk, Andrew R. Poirrette, David W. Rice and Peter Willett.

Krebs Institute, University of Sheffield, Sheffield S10 2TN, United Kingdom.

Abstract

The use of graph-theoretical algorithms in the similarity searching of databases of 3-dimensional structures of protein molecules is discussed. Two levels of searching are considered: (1) comparison of protein structures at the level of alpha helices and beta strands in 3-dimensional space; and (2) searching for three-dimensional patterns of sidechain functional groups.

Introduction.

The last five to ten years have seen rapid developments in database techniques for examining and comparing macromolecular structures. The ability to make such comparisons is of great importance in molecular biology, biochemistry and biotechnology. Early work in this area concentrated on the detection of similarities in the one-dimensional sequences of biological macromolecules, but three-dimensional methods have now become important with the recent rapid increase in the number of known three-dimensional macromolecular structures, specifically of proteins. Some three-dimensional structure comparison methods have built on the earlier one-dimensional ones. However the topic of this chapter will be our development of graph-theory based methods for the detection of molecular similarity in protein molecules. These methods are a direct development of long-established techniques for the representation and searching of small molecule structures (Ash *et al.* 1991).

The repository for macromolecular structures is the Protein Data Bank (PDB), which was established at the Brookhaven National Laboratory in 1971 (Bernstein *et al.* 1977, Abola *et al.* 1987), and which functions as the internationally recognized archive of the three-dimensional structures of biological macromolecules. However, the PDB is not itself a database: it simply consists of more than two thousand ASCII files each of which contains the coordinates of a macromolecule. In recent years there has been very intensive activity amongst researchers in the field of biological structure to derive methodologies to systematize and to enhance understanding of the data that the PDB contains. This is because the understanding of how the amino acid sequence of a protein governs the final 3-dimensional structure, and therefore the function, of that protein, remains the greatest unsolved problem in structural molecular biology (Richards 1991).

It is possible at present to identify two main levels at which molecular similarity is of importance in proteins:

(1) Detection of large-scale similarities between different protein structures, ie: similarities in the way that the linear polypeptide sequence is folded up to form a three-dimensional structure.

(2) Comparative analyses of local aspects of protein structure, for example the examination of specific binding sites, or of the environments of particular sidechains.

Comparison of protein folds in three dimensions.

The fold of a protein is the way in which the regions of helix, strand and random-coil structure within its polypeptide chain are arranged in three dimensions to form its tertiary structure. This is the simplest level at which the three-dimensional structures of different proteins can be compared with one another. Yet, as is indicated below, such similarities may be indicators of remote evolutionary relationships, give clues to functional analogies, or insights into the processes of protein folding.

The comparison of protein folds can be difficult: the three-dimensional structures are frequently complicated, and quite major differences can exist between structures that are, on the basis of sequence similarity, clearly related in evolutionary terms, especially if the similarity is only a partial one. On the other hand structures may sometimes resemble each other very closely, but fail to display any sequence similarity: the classic example of the latter is the "parallel beta barrel" structure which has now been found in more than twenty proteins with no amino-acid sequence homology (Chothia 1988). The interpretation of the meaning of a similarity can therefore be less than straightforward: it may indicate that the proteins are evolutionarily related ("divergent evolution"); that they are unrelated but have evolved similar structures because they carry out similar functions ("convergent evolution"); or that the common structure is simply a particularly stable one that is adopted by a large number of proteins.

Previous work in the field of similarity searching in proteins has led to the development of techniques to carry out numerous comparative tasks in molecular biology, ranging from the creation of sequence alignments to the comparison of 3-D protein structures in geometric terms. In the latter area, much of the work to date has concentrated on the detection of similarity in folding patterns. patterns. The main emphasis has been on the comparison of protein folds by alignment of large portions of protein structures to locate maximal lengths of superimposable main chain. Rao and Rossmann (1977) noted that similar three-dimensional arrangements of alpha helices and beta strands could occur in different protein structures. They called these arrangements "super-secondary structures", and they are also now known as folding "motifs". The earliest quantitative methods for protein structure comparisons were developed by Rossmann and Argos (1977) and by Remington and Matthews (1978). Both methods require considerable computing time to compare a pair of proteins and are consequently not suitable for conducting rapid searches for structural similarities between a protein or a motif and all the other proteins in the PDB. The rapid expansion in the number of known protein structures during the 1980s led to a resurgence of interest in this area, and there has been intense research activity in this field over the last five years. Thus, more recent approaches to the problem of comparing folds include those reported by Taylor and Orengo (1989) who extended the dynamic programming techniques of Needleman & Wunsch (1970); Sali and Blundell (1990) combined dynamic programming with simulated annealing; and Vriend and Sander (1991) who have both developed methods by which proteins may be compared by clustering together similar substructures. At the same time as these developments, comparison methods for proteins based on graph theory were developed in our laboratories (Mitchell *et al.* 1990; Grindley *et al.* 1993) and elsewhere (Subbarao and Haneef, 1991).

Use of graph theoretical methods for detection of similarity in protein folds.

For several years we have been involved in a wide-ranging project to develop methods for the representation and searching of the three-dimensional (3-D) protein structures in the Brookhaven Protein Data Bank. Our work derives from the graph-theoretic methods that are used for the storage and retrieval of information pertaining to both two-dimensional (2-D) and 3-D small molecules (eg: Ash *et al.* 1991).

A molecule in a chemical information system is represented by a labelled graph, in which the nodes and edges represent the atoms and bonds, respectively, of a 2-D molecule, i.e., a planar chemical structure diagram, or the atoms and inter-atomic distances, respectively, of a 3-D molecule. This graph-theoretic representation enables searching operations on databases of chemical structures to be implemented using isomorphism algorithms, which compare one graph with another to determine the structural relationships that exist between them. Thus, subgraph isomorphism algorithms are used in substructure searching systems, which retrieve all molecules from a database that contain a user-defined partial structure, eg: all molecules containing a quinazoline ring system.

We have applied these graph theoretic techniques to protein structure comparison. We also made one other important simplification of the problem: the majority of the comparison methods outlined in the previous section used protein alpha-carbon coordinates for their searches. However, we have reduced the problem by representing alpha helices and beta strands as straight line segments. A protein can then be represented by a matrix containing information on the relative positions and orientations of pairs of these linear segments. Comparisons between proteins can be effected by comparing these matrices.

The graph representations of protein folds are constructed so that the nodes of the graph are the secondary structure elements (ie: the alpha helices and beta strands) of each protein, whilst the edges are the distances and angles between them. In order to accomplish this, regions of helix and strand in proteins in the Protein Data Bank are assigned using the algorithm of Kabsch and Sander (1983). The position and direction of each secondary structure element (SSE) is then approximated by a vector in 3-dimensional space which corresponds to the axis of an idealized helix or strand superposed on the real helix or strand by least squares. The torsional angles, closest approach distances and distances between midpoints of each pair of SSEs within each protein in the PDB are stored in a database as a labelled graph. The nodes of the graph are the linear representations of the SSEs, and the edges of the graph the distances and angles between them (Mitchell *et al.,* 1990).

These earlier small molecule studies involved a detailed comparison of a range of different subgraph and maximal common subgraph isomorphism algorithms and concluded that the algorithms due to Ullmann (1977) and to Bron and Kerbosch (1973) were the most efficient of those that were tested. Two programs, POSSUM (Mitchell *et al.,* 1990) and PROTEP (Grindley *et al.*, 1993) implement the Ullmann and the Bron and Kerbosch algorithms respectively. Later, we describe the ASSAM program which is used for more detailed sidechain-level searching: currently this program implements just the Ullmann algorithm, although the application of the Bron & Kerbosch algorithm to the processing of sidechain data is under investigation in our laboratories at present.

(i) Searching for folding motifs: the POSSUM program.

Our first program, POSSUM (Mitchell *et al.*, 1990), used the subgraph isomorphism algorithm of Ullmann (1976) to identify all protein structures that contained a user-defined structure motif described in terms of alpha helices and beta strands (Secondary Structure Elements or SSEs) and the distances and angles between them. Using POSSUM we were, for example, able to detect a previously unrecognised and intriguing structural homology between the CheY bacterial signal transduction protein and EFTu, an elongation factor related to G proteins (Artymiuk *et al.*, 1990). The program also proved very useful in conducting a wide-ranging survey of beta topologies and psi loops (Ujah, 1992; Artymiuk *et al.*, 1994a), which showed that very few of the possible types of sheet were found to occur in practice, and also allowed the identification of several previously unidentified instances of psi loops.

(ii) Maximal common subgraph searching: the PROTEP program.

The subgraph isomorphism algorithm used in POSSUM requires all the SSEs in the pattern to be present in a structure in the PDB, if that structure is to be retrieved; partial matches between pattern and structure cannot be found. Over the last few years, we have developed a new program, PROTEP (Grindley *et al.*, 1993), in which the same graph representations of proteins are searched using a maximal common subgraph algorithm. Although computationally more intensive than POSSUM, the program is still fast and a search of the entire PDB typically requires 10-30 minutes on an R4000 SG Indigo workstation. PROTEP is very powerful and flexible, because it allows the rapid location of any structural overlaps, whether partial or complete, between a query structure and any of the other proteins in the PDB.

PROTEP has been used to identify several previously unrecognised 3-D resemblances between families of proteins with no obvious sequence homology. In some cases these similarities may indicate a very distant common ancestry, as in the striking similarity we established between the families of Zn^{2+} aminopeptidases and carboxypeptidases (Artymiuk *et al.*, 1992); or between the ribonuclease H and connection domains of reverse transcriptase, and between both these domains and the ATPase fold (Artymiuk *et al.*, 1993). In other cases the similarity is more likely to represent convergent evolution towards a stable structure: an example of this is the structural resemblance we have shown between the core beta sheets of the enzyme protocatechuate 3,4-dioxygenase and two other completely unrelated proteins, the intensely sweet protein thaumatin, and the thyroid-hormone transporting protein prealbumin (Artymiuk *et al.*, 1994a).

One example of a similarity that we have found which may well be indicative of an evolutionary relationship involves the structure of serine t-RNA synthetase and biotin repressor

(Artymiuk *et al.*, 1994b). SerRS was the first structure of a Class II aminoacyl t-RNA synthetase to be solved (Cusack *et al.*, 1990). A PROTEP search revealed a strong resemblance between the core beta sheets of the catalytic domains of SerRS and of BirA, the biotin synthetase/repressor, solved in Brian Matthews group (Wilson *et al.*, 1992). This previously unrecognised similarity (see Figures 1), comprises seven beta strands with identical topology and sequence order, as well two helices on either side of the sheet which are also topologically equivalent and occur in the same sequence order, although they superpose less well in three dimensions. A further search showed that no other proteins in the PDB contain this seven-stranded beta motif, which had been identified as a novel fold when the BirA structure was solved (Wilson *et al.*, 1992). The active sites of the two proteins are very similarly positioned in the catalytic domains, and sidechains from analogous loops in the structure participate in substrate binding. There are obvious analogies in the reactions catalysed by the two enzymes (both proceed via acyl adenylate intermediates), which strengthen the probability of a remote common evolutionary ancestor. However, there are no detectable sequence similarities, and, apart from the similarity in the catalytic domains, the two enzymes are otherwise very different in structure. This is perhaps unsurprising because, after the initial creation of the acyl intermediate, the subsequent transfer of the acyl group is to very different molecules: t-RNA^{Ser} in the case of SerRS, and to a lysine on the biotin carboxyl carrier protein in the case of BirA. Nevertheless the finding is interesting because it may be indicative of the possible evolution of the Class II t-RNA synthetases (Cusack, 1994).

Searching for clusters of sidechains using the ASSAM program.

The ASSAM program, extends the above methods to the representation and searching of arrangements of protein sidechains in 3-D space (Artymiuk *et al.*, 1994c). In ASSAM, the nodes of the graph are the atoms in the sidechains, and the edges are the inter-atomic distances. In principle, all of the atoms in each sidechain could be used in the representation, but in practice it became clear that ambiguities of sidechain positioning due to variation or disorder could be accommodated by using a simplified representation of the sidechain positions, in which each sidechain is characterised by a small number of pseudo-atoms. This approach not only enhances the flexibility of the query procedure, but also improves the speed of searching because of the smaller numbers of nodes that need to be considered in the graph-matching operations when compared with a representation involving all of the atoms in a sidechain. The latter is of particular importance in the context of a database-searching program, since a subgraph isomorphism algorithm must, of necessity, have a running time that is proportional to the factorials of the numbers of nodes in the graphs that are being matched. We therefore define the relative orientations of sidechains in space by means of distances (the edges of the graphs) between

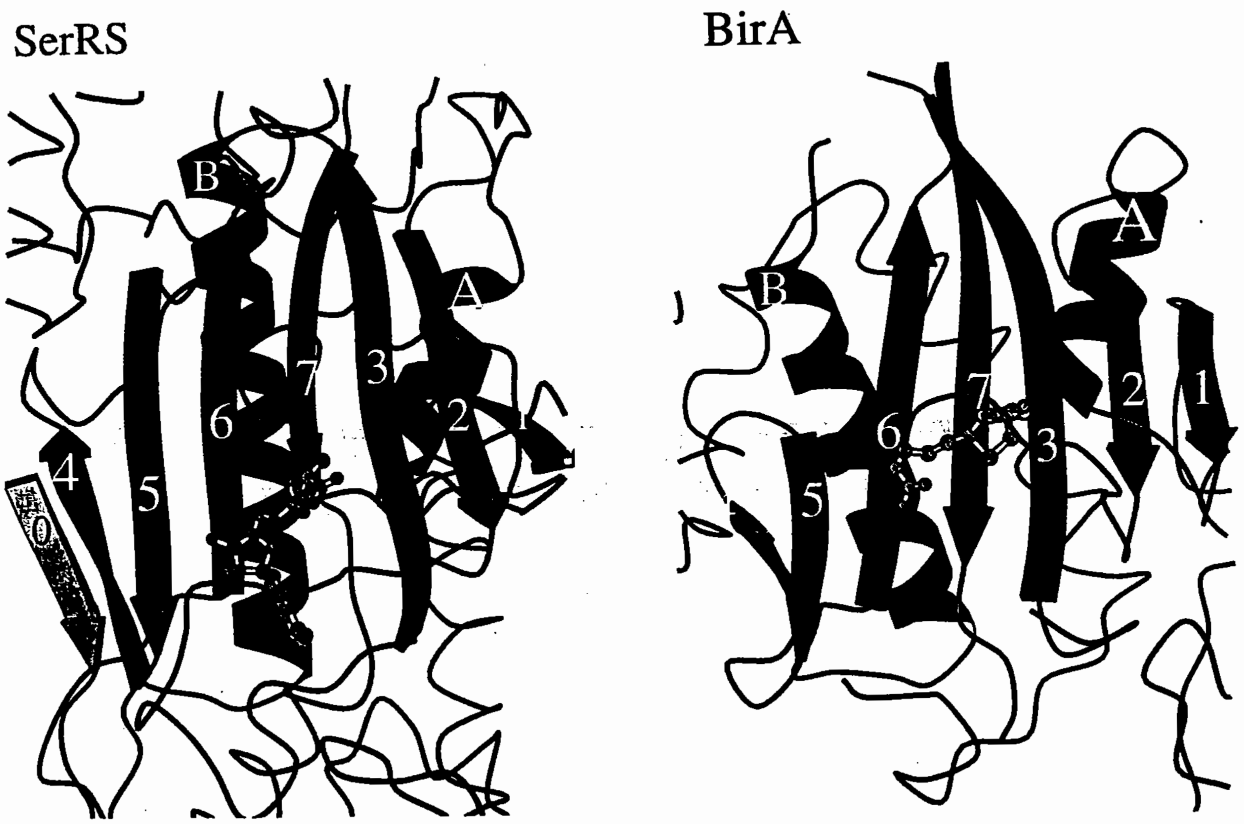


Figure 1 Chain traces, produced using Molscript (Kraulis, 1991) showing the catalytic domains of SerRS (left) and BirA (right). Equivalenced beta strands are shown as sequentially numbered arrows. The two helices A and B also occur in the correct sequence order. In SerRS there is an additional strand marked "0" from the N-terminal domain. A ball-and-stick representation is used to show the binding sites of seryl-hydoxamate-AMP in SerRS, and biotin in BirA, respectively

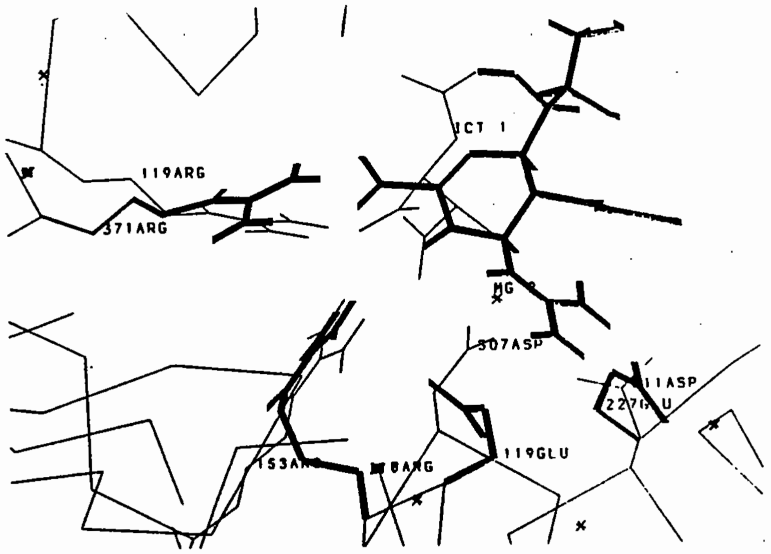


Figure 2 Superposition of the sialidase and ICDH binding sites. Arginines 371 and 118 and glutamates 119 and 227 from sialidase are shown in bold lines, as is the inhibitor 4-guanidino-Neu5Ac2en (upper right). Arginines 119 and 153 and aspartates 307 and 311 of ICDH, together with isocitrate and magnesium ion, are shown superposed on the sialidase in light lines.

pseudo-atoms representing the sidechain (the nodes of the graph). In practice, each sidechain is represented by just two pseudo-atoms, one near to the start and the other near to the end of the functional part of the sidechain: we shall refer to these as the "S" and "E" pseudo-atoms, respectively. In addition a midpoint pseudo atom, "M", is sometimes used.

A search pattern is defined by specifying a set of residues from the coordinate set of a protein the PDB, and the search matrix containing the query vectors and inter-vector distances is then generated automatically. The user is prompted for the tolerances that are to be applied to each of the query distances, and for any generic amino acid types that are to be employed in the search. Each PDB structure in turn is read in from disk, and then the matrix for the appropriate amino acids calculated as described above. The query matrix is sought in the protein matrix using the Ullmann algorithm (1976), and hits output for subsequent inspection. The program has been implemented in FORTRAN 77 on Silicon Graphics and Evans and Sutherland workstations running under the UNIX operating system, and on DEC Alphas running openVMS.

To test the program, patterns were constructed using commonly occurring or well-characterized clusters of residues, so that it would be possible to compare the matches identified by the program with those one would expect to find. A pattern constructed from the Asp-His-Ser catalytic triad residues from alpha-chymotrypsin was used to test the program in this way: in addition to correctly retrieving the known instances of catalytic triad the program detected a surprising and intriguing instance of a second Asp-His-Ser triad in certain chymotrypsinogen and trypsinogen structures (Artymiuk *et al.*, 1994c). This search was carried out on a 791-structure subset of the April 1993 release of the PDB and required less than 3 minutes of real and CPU time on a Silicon Graphics R4000 Indigo workstation.

A recent example of the use of the ASSAM programme is the detection of a structural similarity between binding sites in influenza sialidase and isocitrate dehydrogenase (Poirrette *et al.*, 1994.)

Recently, von Itzstein *et al.* (1993) described the design of potential anti-influenza drugs which operate by binding to a specific site on the influenza sialidase molecule, thereby inhibiting it. A key feature of their analysis of the sialidase structure was the recognition of a negatively charged patch on the sialidase surface provided by two glutamate residues and not used in the binding of sialic acid. This patch was exploited in the rational design process by the synthesis of a sialic acid analogue bearing an additional positive charge from either an amino- or a guanidinyl-group appropriately placed to interact with this region of the sialidase surface. This work constituted a long-awaited example of rational computer-assisted design of a new drug based on the crystal structure of a target protein, and highlighted the great importance of knowing the three-dimensional structures of medically important macromolecules.

Von Itzstein *et al.* identified a group of four sidechains as being involved in drug-binding in the influenza sialidase (two arginines, 371 and 118, and two carboxylic acid groups, glutamates 119 and 227). We were interested to see whether any other proteins contained a similar constellation of residues, and so coordinates for these sidechains were taken from the sialidase structure (Varghese *et al.*, 1991) and were used to generate a search pattern corresponding to this sidechain cluster. Using the ASSAM program we compared this three-dimensional arrangement of sidechains with all the other protein structures in the Protein Data Bank (PDB (October 1992 release). The search revealed (Poirrette *et al.*, 1994) that there is a very similar cluster of four sidechains in the binding pocket for the isocitrate/Mg^{2+} complex in the active site of *E.coli* isocitrate dehydrogenase (ICDH, (Hurley *et al.*, 1990), an enzyme which catalyses the $NADP^{+}$-linked oxidative decarboxylation of isocitrate to 2-oxoglutarate.

The two sets of sidechains from sialidase and from ICDH are shown superposed in Figure 2 where it can be seen that in ICDH the positions of Arginines 119 and 153 correspond closely to the two Arginines 371 and 118 respectively in the sialidase, with excellent overlap of the guanidinium groups. In ICDH these two arginines interact with two carboxyl groups of the isocitrate, one of which, the C3 carboxyl, occupies an equivalent position to the carboxyl group of the inhibitor 4-guanidino-Neu5Ac2en in sialidase. The two carboxyl groups in the ICDH site are provided by two aspartates (aspartates 307 and 311) which are slightly displaced with respect to their counterparts in sialidase (glutamates 119 and 227 respectively). In sialidase these carboxyls coordinate the positive charge of the 4-guanidinyl group of the inhibitor, and in ICDH the equivalent carboxyls analogously coordinate the positive charge of the Mg^{2+} ion in the complex.

Clearly, this result does not mean that the sialidase would necessarily be expected to bind isocitrate. The binding site for sialic acid comprises many more residues than the two arginines and two aspartates that were highlighted by von Itzstein *et al.* (1993) and which we used as our search pattern. In fact these four residues constitute the only major area of resemblance between the binding sites in the two proteins; the folds of the polypeptide chain, and the positions of other residues in the active site, are quite different. Nevertheless, this finding does suggest that compounds with an isocitrate-like framework and a similar pattern of charge distribution might form a suitable alternative starting point for the design of new families of anti-influenza drugs.

Conclusions.

An important factor in the development of molecular biology has been the availability of computational tools for the detection of similarities in the one-dimensional sequences of proteins and nucleic acids (Lesk, 1988). It is similarly of the greatest importance to be able to detect structural analogies within the rapidly growing database of three-dimensional protein and nucleic acid structures, in order to enhance understanding of structure/function relationships in biological macromolecules. The folding pattern similarities revealed by PROTEP, and the influenza sialidase binding site example discussed in the previous section, suggest that such comparative studies may give valuable insights into structures of medical interest, and represent another valuable approach to the rational design of novel inhibitors.

In more general terms, it is clear that structural comparisons of the kind reported here may be more widely applicable. At present the database of known three-dimensional protein structures is very small, consisting of only a few hundred distinct structures. As this database expands, it is increasingly likely that more similarities of the kind we have described above will be observed between otherwise disparate proteins. At the scientific level, such similarities will enhance our understanding of structure/function relationships in proteins; whilst technologically they may prove valuable both in protein engineering and in the search for new lead compounds in drug design. Thus it seems likely that structural comparisons will be a useful weapon to be used in conjunction with other modelling procedures.

Acknowledgements.

We thank the Royal Society, MRC and Tripos Associates Inc. for support.

References

Ash JE, Warr WA, Willett P (eds.) (1991) Chemical Structure Systems, Ellis Horwood, Chichester

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jnr., Brice MD, Rodgers JR, Kennard O, Shimanouchi M, Tasumi M (1977) J. Molec. Biol. **112,** 535.

Abola EE, Bernstein FC, Bryant SH, Koetzle TF, Weng J (1987) in, Allen FH , Bergeroff G, Sievers R (eds.) Crystallographic Databases - Information Content, Software Systems, Scientific Applications, pp 107-132, Data Commission of the International Union of

Crystallography, Bonn/Cambridge/Chester.

Richards FM (1991) Scientific American **264,** 54.

Chothia C (1988) Nature **333**, 598

Rao ST, Rossmann MG (1973) J Molec. Biol. **76**, 241

Rossmann MG, Argos P (1976) J. Molec. Biol. **105**, 75

Remington SJ, Matthews BW (1978) Proc. Nat. Acad. Sci. U.S.A. 75, 2180

Taylor WR, Orengo CA (1989) J. Molec. Biol. **208**, 1

Needleman SB, Wunsch CD (1970) J. Molec. Biol. **48**, 443

Sali A, Blundell TL (1990) J Molec. Biol. **212**, 403

Vriend G, Sander C (1991) Proteins, Struct. Funct. and Genet. **11**, 52

Mitchell EM, Artymiuk PJ, Rice DW, Willett P (1990) J. Molec. Biol. **212**, 151

Grindley HM, Artymiuk PJ, Rice DW, Willett P (1993) J. Molec. Biol. **229**, 707

Subbarao N, Haneef I (1991) Prot. Eng. **4**, 877

Kabsch W, Sander C (1983) Biopolymers **22**, 2577

Ullmann JR (1976) J.Ass.Comp.Mach. **16**, 31

Bron C & Kerbosch J (1973) Comm. Ass.Comp.Mach. **16**, 575

Artymiuk PJ, Rice DW, Mitchell EM, Willett P (1990) Prot. Eng. **4**, 39

Ujah EC (1992) PhD Thesis, University of Sheffield, Sheffield

Artymiuk PJ, Grindley HM, Poirrette AR, Rice DW, Ujah EC, Willett P (1994a) J. Chem. Inf. Comput. Sci. **34**, 54

Artymiuk PJ, Grindley HM, Park JE, Rice DW, Willett P (1992) FEBS Lett. **303**, 48

Artymiuk PJ, Grindley HM, Kumar K, Rice DW, Willett P (1993) FEBS Lett. **324**, 15

Artymiuk PJ, Grindley HM, MacKenzie AB, Rice DW, Ujah EC & Willett P (in the press) In "Molecular Similarity and Reactivity, from quantum chemical to phenomenological approach" (ed. R. Carbo) Kluwer Academic Publishers.

Artymiuk PJ, Rice DW, Poirrette AR & Willett P (1994b) Nature Struct. Biol. **1**, 758

Cusack S, Berthet-Colominas C, Härtlein M, Nassar N & Lebermann R (1990) Nature, **347**, 249

Wilson KP, Shewchuk LM, Brennan RG, Otsuka AJ & Matthews BW (1992) Proc. natl. Acad. Sci. USA **89**, 9257.

Cusack S (1994) Nature Struct. Biol. **1**, 760

Kraulis PA (1991) J.Appl.Cryst. **24**, 946

Artymiuk PJ, Poirrette AR, Grindley HM, Rice DW & Willett P (1994c) J. Molec. Biol. **243**, 327

Poirrette AR, Artymiuk PJ, Grindley HM, Rice DW & Willett P (1994) Protein Science **3**, 1128

von Itzstein M, Wu W-Y, Kok GB, Pegg MS, Dyason JC, Jin B, Phan TV, Smythe ML, White HF, Oliver SW, Colman PM, Varghese JN, Ryan DM, Woods JM, Bethell RC, Hotham VJ, Cameron JM, Penn CR (1993) Nature **363**, 418

Varghese JN, Colman PM (1991) J. Molec. Biol. **221**, 473

Hurley JH, Dean AM, Sohl JL, Koshland DE Jr, Stroud RM (1990) Science **249**, 1012

Lesk AM (ed.) (1988) Computational Molecular Biology Oxford University Press, Oxford

Software tools for protein structure analysis and fold classification

A.D. Michie, E.G. Hutchinson, R.A. Laskowski, C.A. Orengo & J.M. Thornton

Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College London, Gower Street, LONDON WC1E 6BT, UK.

Abstract. Some new software tools are described which help the crystallographer make the most of a newly determined protein structure. PROCOMP, based on PROCHECK (Laskowski et al., 1993), compares the local geometry of two identical or closely related structures. This may be used during refinement to highlight changes in a model. PROMOTIF identifies and classifies local structural motifs in proteins (e.g. β-turns or bulges). A newly determined structure may be compared to all others in the database using the SSAP algorithm (Taylor & Orengo, 1989). This algorithm, combined with the commonly accepted description of folds, has been used to classify all protein structures in the Protein Databank into fold families indexed by 'CATH numbers' akin to the E.C. nomenclature for enzymes. Successive numbers represent Class, Architecture, Topology and Homology families. Thus the globins have CATH number 1.1.1.1, whilst immunoglobulins are described by 2.1.1.1 . This classification scheme is accessible on the World Wide Web.

Introduction

The number of protein structures being determined is increasing rapidly. This has had several consequences for the crystallographer solving a new structure. Firstly a new structure should be reviewed in the light of the accumulated structural database and this becomes more difficult as the size of the database increases. A new structure may look like one that has already been seen; this is important to recognise, since the relationship may well have functional and evolutionary implications. Secondly it is necessary to assess the quality and 'normality' of a structure, in comparison to other structures determined at the same resolution. Are there unusual conformations or motifs that may be important for the protein's function? Thirdly an informed description of the new structure should be given in terms of the 'known' structural motifs, (e.g. the different turn types or sheet motifs), that have been described previously in the literature.

A different problem confronts most biologists who wish to use the Protein Structure Databank (PDB) (Bernstein et al., 1977) as a source of information about a single protein or a family of related structures. This is no longer a trivial task. All members of a family must be identified and many families are large and sometimes not obviously related. For example there are more than 120 entries for the serine proteinases in the database, which would need to be inspected and described. How does a non-expert access these structures and compare and contrast them?

Lastly there are the protein structure aficionados, who wish to browse through the database structure-gazing! As new structures appear in the literature almost every day, manual inspection of all structures becomes increasingly difficult, especially as many of the new structures are extremely large.

Therefore for all these different reasons, it is necessary to devise automated methods to make structure analysis and classification easier. Such software tools will be useful in the initial inspection of a new model by a crystallographer, when the aim is to 'make the most of the model'! They will also be used by the non-expert molecular biologist, with a new interest in protein structures. Below we list some new tools we have developed, which could be used by a crystallographer both during refinement and after a structure has been completed. These tools only require atomic coordinates, and do not require electron density. Their functions are as follows:

1. Comparison of the conformations of two identical or related structures - PROCOMP
2. Identification and characterisation of structural motifs in the protein - PROMOTIF.
3. Comparison of a structure to all others in the PDB and identification of those with similar folds - using the SSAP algorithm to give the CATH classification.

In this paper we shall present a brief description of each tool and its availability.

1. PROCOMP

During our work on protein structure analysis and classification, we have found that we often need to compare related structures, looking for similarities or differences in their conformations. The differences may represent real changes in conformation (e.g. when a ligand binds to a protein) or perhaps uncertainties in structure determination.

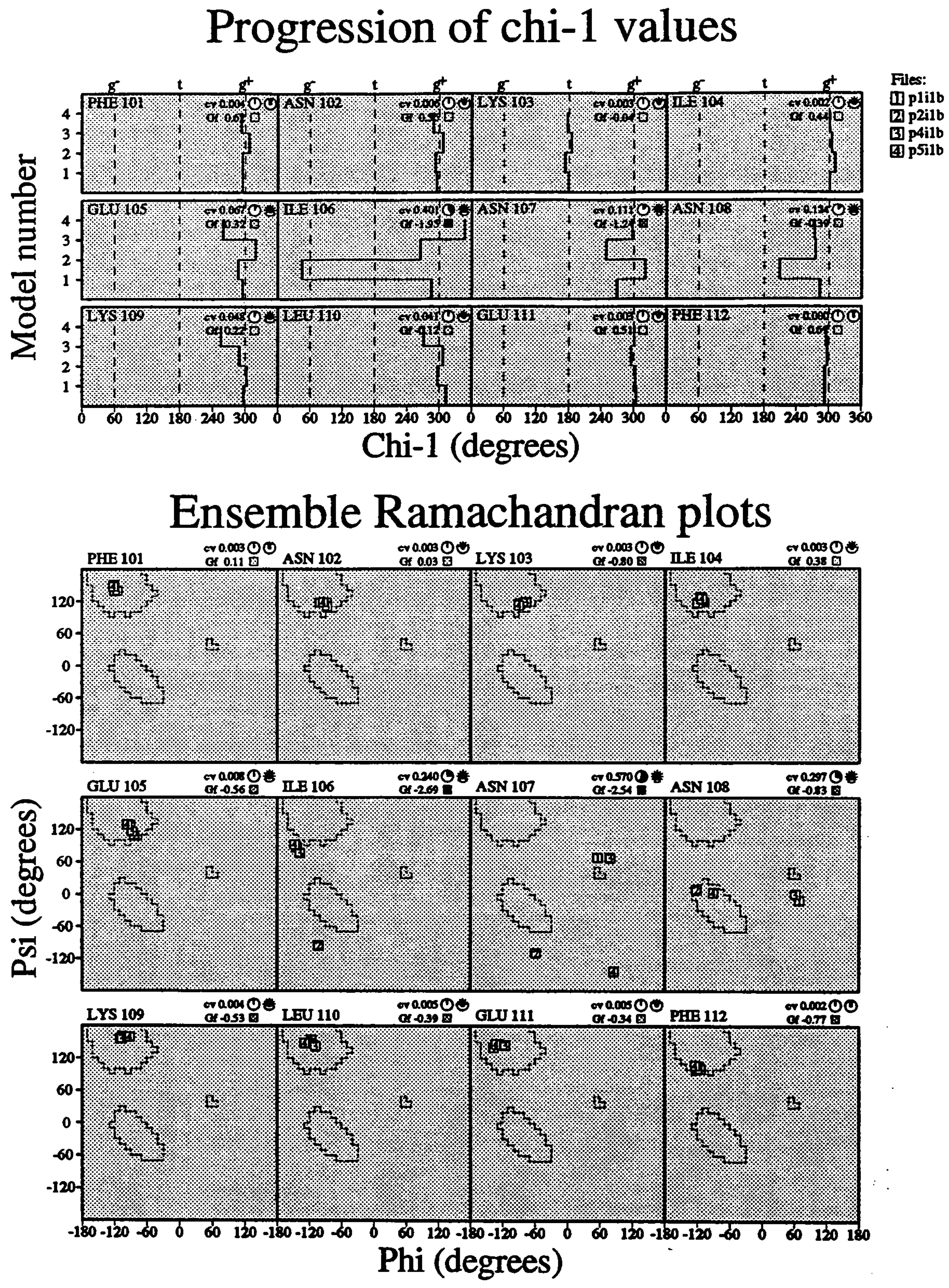
PROCOMP is a set of programs, related to PROCHECK (Laskowski et al., 1993), which compares the overall and residue-by-residue geometry of a set of closely-related structures. The structures might be separate members of a family of proteins, or models of the same structure saved during different stages of refinement, or even a homology model and the structure on which it was based.

The program outputs a number of PostScript files showing the comparisons, including:-

a. Ramachandran plot for all residues
b. Ramachandran plot for Gly & Pro only
c. χ_1-χ_2 plots
d. Dihedral angles, compared on a residue-by-residue basis:
    Φ, ψ, χ_1, χ_2 torsion angles; Ramachandran and χ_1-χ_2 plots
e. Residue properties, including:
    Absolute ω torsion angle, χ_1 deviations; average secondary structure; circular variances of the dihedral angles; RMS deviations for main-chain and side-chain atoms; residue-by-residue G-factors (geometry factors representing deviations from 'normality')
f. Secondary structure in each model

Figure 1 shows an example of the residue-by-residue χ_1 torsion angles and Ramachandran plots for four independently-solved structures of human interleukin 1-β (PDB codes 1i1b (Finzel et al., 1990), 2i1b (Priestle et al., 1989), 4i1b (Veerapandian et al., 1992), 5i1b (Ohlendorf et al., (unpub.)).

The program is freely available for academic users on the ftp server (128.40.46.11) in the pub/procheck directory. Alternatively, contact Roman Laskowski via e-mail at the following address: **roman@biochem.ucl.ac.uk** . Sample output can be viewed via World



cv = Circular Variance (low values signify high clustering of the data points). ✹ Accessible Ⓤ Buried

Gf = Average G-factor for the residue (the higher the value the more favourable the conformations) based on analysis of high-res. Xstal structures

Figure 1

Wide Web on: **http://www.biochem.ucl.ac.uk/bsm/**

2. PROMOTIF

PROMOTIF is a suite of programs which analyses a protein coordinate file and provides details of the structural motifs in the protein. Currently the program provides information about the following structural features: secondary structure; β- and γ-turns; helical geometry and interactions; β-strands and β-sheet topology; β-bulges; β-hairpins; β-α-β units and ψ-loops; and main-chain hydrogen bonding patterns.

The program takes as input a Brookhaven format PDB file and produces a series of output files for each motif. These include flat files, black and white PostScript tables and black-and-white and colour PostScript schematic diagrams.

Most of the motifs are identified and classified according to rules defined in published papers. The secondary structure of the protein is calculated using an algorithm which is essentially the same as the standard DSSP algorithm of Kabsch and Sander (1983). (The secondary structure differs slightly in that one extra residue is added to the beginning and end of each secondary structural unit where possible to conform more closely to the IUPAC standard). This secondary structure file provides the raw data used for the remainder of the analyses.

Some sample outputs from the program are discussed below.

β-Turns

A β-turn is defined for 4 consecutive residues (i to i+3) if the $C_\alpha(i)$-$C_\alpha(i+3)$ distance is less than 7 Å and if the central two residues are not helical (Lewis et al., 1973). The turns are classified according to the φ, ψ angles of residues i+1 and i+2 (Venkatachalam, 1968; Richardson, 1981). A more recent analysis of turns is described in Hutchinson & Thornton (1994). Figure 2 shows part of the PROMOTIF output for the β-turns in carboxypeptidase A (5CPA) (Rees et al., 1983). This and the following three figures are produced in colour by PROMOTIF. A table (not shown) provides more precise data on the φ,ψ angles for each turn.

Similar data are produced for γ-turns, which involve 3 residues: i, i+1 and i+2 with a hydrogen bond between residues i and i+2 and the φ,ψ angles of residue i+1 falling within 40° of the standard angles of either of the two classes: Classic (φ,ψ = 75°, -64°) and Inverse (φ,ψ = -79°, 69°). (Rose et al., 1985)

β-Bulges

β-bulges are regions of irregularity in a β-sheet, where the normal pattern of hydrogen bonding is disrupted, e.g. by the insertion of an extra residue. PROMOTIF uses the algorithm recently described by Chan et al (1993) to identify and classify bulges in proteins. Figure 3 shows the bulges found in carboxypeptidase A. Bulges are classed as Parallel or Antiparallel depending on the relative orientation of the two strands involved. The bulges are further subdivided into *classic*, *wide*, *bent*, *G1* and *special* types depending on the number of residues involved and the hydrogen bonding pattern. Classic bulges (shown here) involve an extra residue on one strand relative to its neighbouring strand. The hydrogen bonding patterns in parallel and antiparallel cases are different. G1 bulges occur only in antiparallel sheets; in

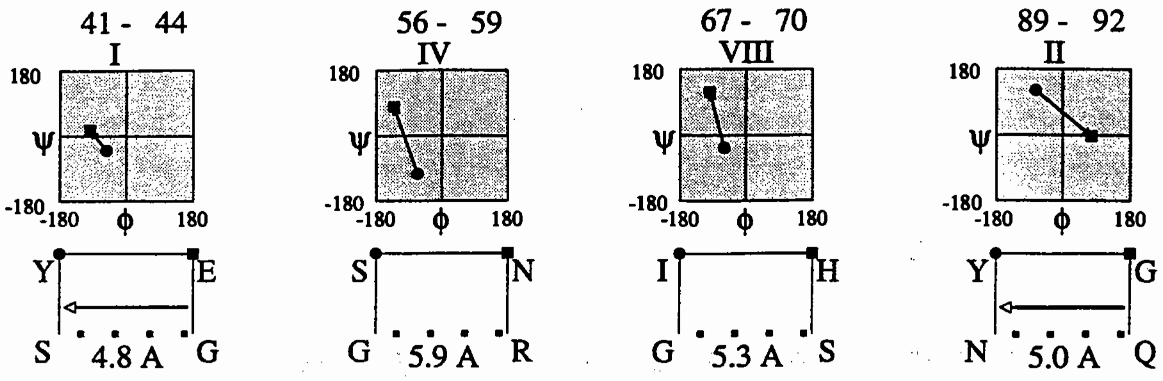


Figure 2. Some of the beta turns in carboxypeptidase A (5CPA)

For each turn, the residue numbes involved and the turn type are shown. A Ramachandran plot shows the phi/psi values of residues i+1 (represented by a circle) and i+2 (square) in the turn, with the arrow from i+1 to i+2. The bottom part of the figure shows a schematic diagram of the turn, with the one-letter code of the residues involved and the C_α(i)-C_α (i+3) distance in Å. An arrow indicates a hydrogen bond between the NH of residue i+3 and the CO of residue (i), where present.

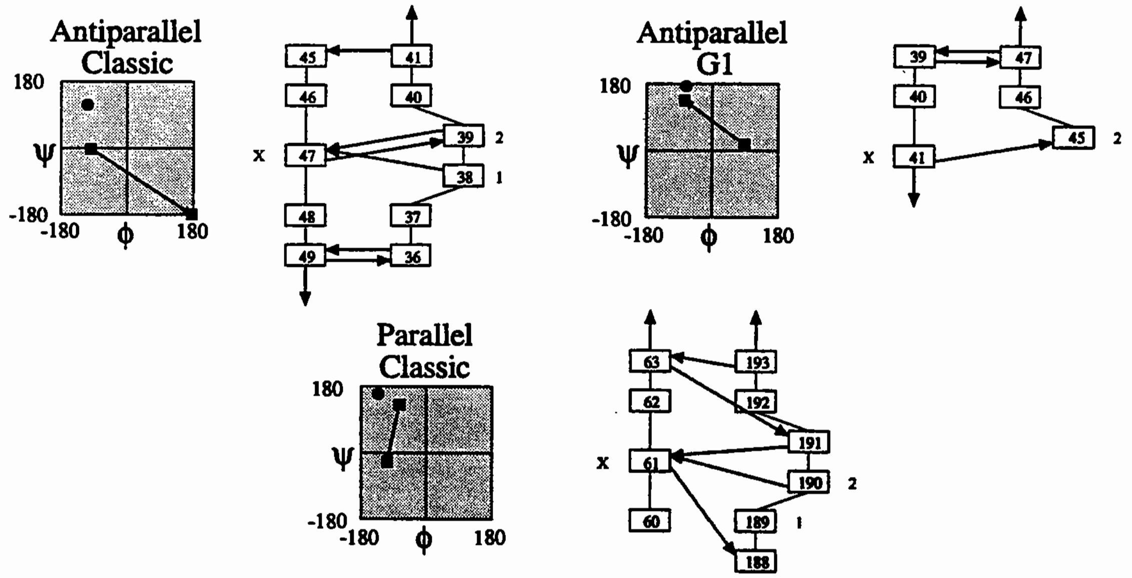
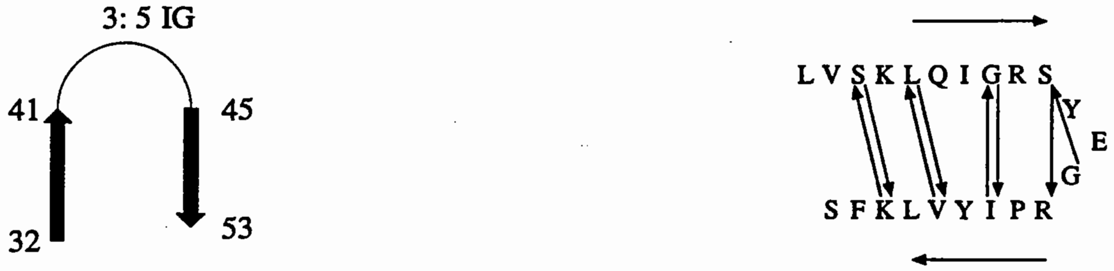


Figure 3. Beta bulges in carboxypeptidase A (5CPA)

A Ramachandran plot for each bulge is plotted showing the phi/psi values of residues X on the normal strand (circle), and 1 and 2 on the bulged strand (square) (For special bulges, in which there is a larger insertion, more values are plotted). To the right of the Ramachandran plot there is a schematic diagram showing the detailed main-chain hydrogen bonding pattern around the bulge.



Strand 1		Strand 2		Number of residues		Hairpin
Start	End	Start	End	Strand 1	Strand 2	Class
32	41	45	53	10	9	3: 5 IG

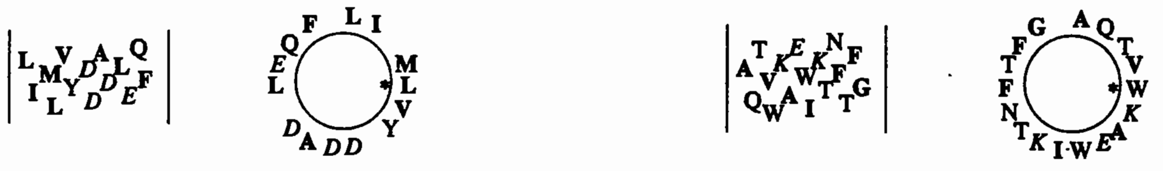
Figure 4. Schematic diagram and table of a beta-hairpin in carboxypeptidase A (5CPA)

The hairpin is type 3:5 IG (where IG indicates this special combination of turn and bulge). The rightmost figure shows the sequence of residues involved in the hairpin, with main chain hydrogen bonds drawn as arrows. The turn and bulge involved in this part of the sequence are shown in Figs. 2 and 3.

Bold: Hydrophobic

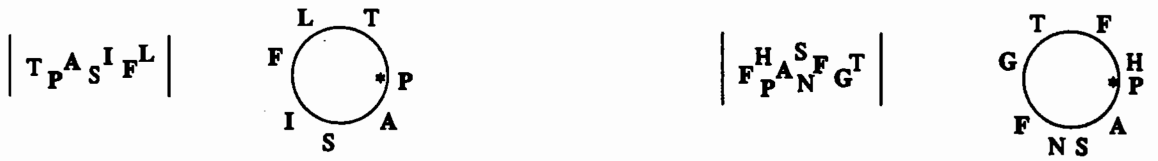
Normal : Polar

Italic : Charged



Helix 1 (15 - 28)
LDEIYDFMDLLVAQ

Helix 2 (73 - 89)
WITQATGVWFAKKFTEN



Helix 3 (94 - 100)
PSFTAIL

Helix 4 (113 - 121)
PNGFAFTHS

Figure 5. Helical wheels and nets for the first four helices in carboxypeptidase A (5CPA)

Asterisks represent the first residues in each wheel.

these cases residue 1 is in the α_L conformation and is therefore usually glycine. This usually occurs at the end of a β-strand.

β-Hairpins

β-hairpins consist of two β-strands which are antiparallel and hydrogen bonded together (connected by at least one bridge). The hairpins are classified as in Sibanda et al. (1989) using two numbers X:Y, which denote the number of residues in the loop defined using two different IUPAC conventions. For the smaller loops the hairpins are dominated by the formation of β-turns. The 3:5 hairpins are dominated by one well defined conformation which can be described as a type I turn followed by a G1 bulge. An example of such a hairpin is found in carboxypeptidase A (Figure 4).

α-Helices

For each helix identified by the secondary structure assignment program, PROMOTIF gives information about the helix type (α or 3_{10}), number of residues and the secondary structures immediately surrounding the helix. A separate table provides geometrical information about each helix such as the length, unit rise, number of residues per turn, pitch and a measure of linearity. Helical wheels and nets are drawn for each helix, assuming that there are 3.6 residues/turn (Figure 5). The residues in these plots can be colour coded for hydrophobic, polar and charged amino acid types. The program also provides details of the interacting pairs of helices in the protein, their distance of closest approach and the omega angle between the helices.

β-Strands and *β*-sheets

More basic data are provided for each β-strand - the start and end residues, number of residues, the amino acid sequence and the β-sheet to which the strand belongs. For each β-sheet a separate table gives the number of strands, the nature of the sheet (antiparallel, parallel or mixed) and the topology of the sheet, using the nomenclature of Richardson (1981). This assigns a number to the connection between each pair of sequential strands in the sheet. This number represents the number of strands the connection traverses in the sheet, and in which direction, with an 'X' added for crossover connections (parallel strands). The location and number of residues in all β-α-β units and ψ-loops present in the protein are also recorded in separate tables.

Main-chain hydrogen bonding patterns

PROMOTIF will also plot a schematic diagram drawn by the program HERA (Hutchinson and Thornton, 1990) to illustrate the main-chain hydrogen bonding patterns in the β-sheets and α-helices of the protein.

Availability

The program is freely available for academic users. The relevant files can be copied from our anonymous ftp server (address: 128.40.46.11). The files are in the /pub/promotif directory. Alternatively, contact the author (Gail Hutchinson) via e-mail at the following address: **gail@bsm.bioc.ucl.ac.uk** . Sample output can be viewed via World Wide Web on **http://www.biochem.ucl.ac.uk/bsm/** .

3.Fold Analysis

Inspection of the ~3000 protein structures currently in the PDB, shows that many sequences - even some with insignificant sequence similarities - fold in a similar way. Sequences that have >30% identity with each other almost certainly have the same fold, and sequence alignments will suffice to put them into structural families, but to detect similarities between proteins with lower sequence identities, direct structural comparison methods have to be applied. In our laboratory structural comparisons are performed using the program SSAP (Taylor & Orengo, 1989, Orengo et al., 1992), which compares residue 'structural environments', returning a normalised similarity score (0-100) between two folds. Structure pairs that have a sufficiently high SSAP score (and a significant proportion of the larger fold equivalenced) are merged into families.

The CATH Classification

The classification used is based on both sequence and structural comparisons. Proteins are grouped initially according to sequence identity, then single representatives of these sequence families are chosen and clustered by structural similarity using SSAP. The representatives of these clusters are then classified manually into more general groups such as 'class' and 'architecture' - e.g. the 'mainly-α class' and the 'orthogonal architecture'. This structure classification scheme naturally results in a hierarchical tree structure outlining the relationships between folds. The resulting protein domain classification can be found in Orengo et al. (1993) - currently only around 800 of the 3000 structures are non-identical, and these are grouped into slightly over 200 single domain fold families. Methods to automatically assign class and architecture are currently being developed. Fold clusters are given names to describe their common structure and the criteria that set them apart from others.

Some other sequence and structural comparison methods used for clustering common folds into families can be found in Pascarella & Argos (1992), Holm et al. (1993), Overington et al. (1993) and Sali & Blundell (1990).

The ever increasing number of structures means that computerised handling of the hierarchy has become necessary so instead of using 'computer-unfriendly' descriptions, we propose a numerical scheme ('CATH numbers') to facilitate automatic indexing and searching of the underlying data. CATH numbers are assigned to fold clusters at the domain level with multidomain structures being decomposed into individual domains which can then be given a specific CATH number. In addition to the numeric classification, CATH contains other data on the structures (Molscript (Kraulis, 1991) , TOPS (Flores et al., 1994), Hera (Hutchinson & Thornton, 1990) and Ligplot (Wallace et al., (in press)) diagrams for example), and relevant sequence and structure comparison matrices can be examined at the lower, quantitative levels.

CATH has been made accessible in a hypertext form over the 'World Wide Web' for use by text-based or graphical browsers. 'SCOP' (Structural Classification of Proteins), a similar resource (Murzin et al., JMB in press; Barton, 1994), is also available over the Web and the addresses for both are given at the end of this section.

'CATH' is an acronym standing for Class, Architecture, Topology and Homology; the major levels in the hierarchy described below.

Class

- Secondary structure type and distribution along the sequence

The first, most general level of the classification, describes the relative content of α-helices and β-sheets in a similar way to that described by Levitt & Chothia (1976). The four most important classes are mainly-α, mainly-β, alternating α/β and α+β with an additional 'multidomain' class for large structures with multiple subunits (see Table 1). These multidomain structures clearly need to be classified by domain, but this is not yet completed.

Table 1: Class

Class #	Class Name	Description
1	mainly-α	predominantly α-helices, few β-strands
2	mainly-β	predominantly β-strands, few α-helices
3	alternating α/β	helices & strands tending to alternate along the sequence
4	α+β	mixed helices & strands
5	multidomain	large structures, divisible into distinct domains
0	unclassified	structures not yet entered into the scheme

Architecture

- Description of the gross arrangement of secondary structures
- Independent of topology

This level distinguishes structures in the same Class with different architectures, but does not distinguish between different topologies (connectivities). For example, in the mainly-α class (C = 1), we have 3 different architectures where A=1 indicates an orthogonal arrangement of helices, A=2 indicates aligned helices etc., as shown in Table 2. The globins, with a typically 'orthogonal' arrangement of helices would thus be placed under CATH number "1.1" as opposed to members of the cytochrome b562 group in which most helices are 'aligned' in a 4α-helical bundle structure which would be assigned "1.2". But a given architecture will contain structures with diverse connectivities which will be distinguished at the next level (Topology).

Table 2: Architectures in the mainly-α class

Arch. #	Arch. Name	Description
1	orthogonal	helices arranged roughly at right-angles
2	aligned	helices arranged in a parallel/antiparallel manner
3	solenoid	helices forming a solenoid structure

Topology

- Topological description is given by reference to previously observed structures and well-known folds
- Similar structures, but may have diverse functions

The structural groups found at the 'T' level are clustered according to their topological connections and number of secondary structures. In our clustering scheme proteins at this level will have pairwise SSAP scores that are >70. Proteins with the same CAT numbers have the same Class, Architecture and Topology but do not necessarily belong to the same sequence superfamily or have the same function.

Homology

- Highly similar structures
- Often correlates with functional homology

Structures grouped by their high SSAP scores (>80), have the same CATH number and this often suggests that they have similar functions and may have evolved from a common ancestor. Using the same example of alpha.orthogonal.globin-like folds, the erythrocruorins, colicins and phycocyanins all have the same CAT number (1.1.1), but are differentiated by their 'H' numbers (1, 2 and 3 respectively).

Additionally, another level is present in the classification which separates proteins with identical CATH numbers into different sequence families.:

Sequence Family

- >30% sequence identity
- High probability of having similar structure/function

Members which are clustered at this level (having the same CATHS number) have sequence identities >30% and as such are expected to have extremely similar structures and functions - e.g. they may be slightly different examples of the same protein from different species belonging to the same sequence superfamily.

Access via World Wide Web (WWW)

In order to allow widespread access to the classification, it has been represented as a interlinked network of hypertext pages that can be viewed remotely from any suitably equipped computer system. The pages allow access to extra information and diagrams (e.g. Molscripts, TOPS, HERA plots etc.) about given folds and their functional groups in the form of downloadable text/PostScript files as well as basic searching and report facilities. Each individual PDB file contained within a sequence family is available direct from the Brookhaven laboratory and links to other related sites such as sequence databases are also planned. Whilst these pages are currently in a preliminary stage of development, it is hoped that extra functionality can be added to allow more complex searches and programs to be run under the control of the remote user.

CATH can be accessed via the URL **http://www.biochem.ucl.ac.uk/bsm/**
SCOP can be accessed via the URL **http://www.bio.cam.ac.uk/scop/**

CONCLUSION

These tools have been developed during the course of our work to understand how the sequence of a protein determines its structure, and how the structure in turn mediates the biological function. Clearly it is most important to relate structure to function and it would be very useful to have an equivalent functional classification for proteins. For enzymes, the E.C. numbers provide a useful starting point, but there is no equivalent for other functional types, eg transcription factors. Developing such a classification scheme will facilitate the analysis of the human genome, and recognition of distant relationships. Correlation between other factors and fold type are also interesting, e.g. cellular location.

These tools are intended to make structural analysis and classification easier and more accurate, and to help throw new light on structural similarities and differences. However, as with all automatic tools, they should be used with cautious intelligence, and if combined with careful visual inspection of a structure, will allow the most to be extracted from a model!

References

Barton, G. TIBS *19* (1994) 554-555
Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F.Jr., Brice, M.D., Rogers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M.. J. Mol. Biol. *112* (1977) 535-542
Chan, A.W.E., Hutchinson, E.G., Harris, D. and Thornton, J.M.. Prot. Sci. 2 (1993) 1574-1590
Finzel, B.C, Clancy, L.L, Holland, D.R., Muchmore, S.W., Watenpaugh, K.D. and Eisenpahr, H.M.. J. Mol. Biol *209* (1990) 779-791
Flores, T.P, Moss, D.S, and Thornton, J.M.. Prot. Eng. *7* (1994) 31-37
Holm, L., Ouzonis, C., Sander, C., Tuparev, G., Vriend, G. (1993). Prot. Sci. *1* 1691-1698
Hutchinson, E.G. and Thornton, J.M.. Prot. Sci. (1994 in press)
Hutchinson, E.G. and Thornton, J.M.. Prot. Str. Fn. Gen. *8* (1990) 203-212
Kabsch, W. and Sander, C.. Biopolymers *22* (1983) 2577-2637
Kraulis, P.J.. J. App. Cryst. *24* (1991) 946-950
Laskowski, R.A., MacArthur,M.W., Moss,D.S. and Thornton, J.M.. J. Appl. Cryst. *26* (1993) 283-291
Levitt, M. and Chothia, C.. Nature, *261* (1976) 552-558
Lewis, P.N., Momany, F.A. and Scheraga, H.A.. Biochem. Biophys. Acta. *303* (1973) 211-229
Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C.. J. Mol. Biol. (in press)
Ohlendorf, D.H., Treharne, A.C, Weber, P.C., Wendoloski, J.J., Salemme, F.R., Lischwe, M. and Newton, R.C. (unpub.)
Orengo, C.A., Brown, N.P., Taylor, W.R.. Prot. Str. Fn. Gen. *14* (1992) 139-167
Orengo, C.A., Flores, T.P., Taylor, W.R. and Thornton, J.M.. Prot. Eng. *6* (1993) 485-500
Overington, J.P., Zhu, Z.Y., Sali, A., Johnson, M.S., Sowdhamini, R., Louie, G., Blundell, T.L.. Biochem. Soc. Transac. *21* (1993) 597-604
Pascarella, S. and Argos, P.. Prot. Eng. *5* (1992) 121-137
Priestle, J.P, Schaer, H.P and Grütter, M.G.. PNAS USA *86* (1989) 9667-9671
Rees, D.C., Lewis, M. and Lipscomb, W.N.. J. Mol. Biol. *168* (1983) 367-387
Richardson, J.S.. Adv. Prot. Chem. *34* (1981) 167-339
Rose, G.D., Gierasch, L.M. and Smith, J.A.. Adv. Prot. Chem. *37* (1985) 1-109
Sali, A. and Blundell, T.L.. J. Mol. Biol. *21* (1990) 404-428
Sibanda, B.L., Blundell, T.L. and Thornton, J.M.. J. Mol. Biol. *206* (1989) 759-777

Taylor, W.R. and Orengo, C.A.. J. Mol. Biol. *208* (1989) 1-22
Veerapandian, B., Gilliland, G.L., Raag, R, Svensson, A.L.. Prot. Str. Fn. Gen. *12* (1992) 10-23
Venkatachalam. Biopolymers *6* (1968) 1425-1436
Wallace, A.C., Laskowski, R.A. and Thornton, J.M.. Prot. Eng. (1995 in press)

Protein stability in the Archaea

Garry Taylor, Rupert Russell, Jamie Rossjohn, Michael Danson & David Hough
School of Biology and Biochemistry, University of Bath, Bath BA2 7AY

Introduction

The Archaea represent a phylogenetically distinct, evolutionary Domain which comprises organisms that all live in extreme environmental conditions (Woese *et al.*, 1990). The various phenotypes are:

- *Thermophiles*: living from 55 to 110°C, the latter extreme being found in deep sea volcanic steam vents.
- *Halophiles*: living in up to 5M NaCl, although they maintain an isotonic intracellular concentration of KCl.
- *Methanogens*: which are strictly anaerobic.
- *Psychrophiles*: pcr analysis of the prokaryotic biomass of Antarctic waters has recently raised the intriguing possibility that there may be psychrophilic Archaea (living around 0°C), although none have yet been cultured (DeLong *et al.*, 1994).

At Bath we have a programme to study the structural basis of protein stability in the Archaea. This involves cloning and sequencing genes encoding enzymes of central metabolism, through to crystallisation and X-Ray structure determination, as well as site-directed mutagenesis to test the features which may confer stability. This paper presents our recent findings for a series of thermophilic forms of citrate synthase, and a comparative analysis with its structure from pig in order to deduce the thermostabilising features.

Citrate synthase

We have chosen citrate synthase as a model protein because: (i) it is present in almost all organisms, catalysing the entry of carbon into the citric acid cycle, (ii) there are around 20 amino acid sequences available across the three kingdoms (Archaea, Bacteria and Eukarya), (iii) there is a good structure at 1.7Å available from pig (Remington *et al.*, 1982), a mesophilic organism (i.e. one which lives at around 20 to 40°C), and (iv) the enzyme from Archaea and Eukarya is dimeric, allowing analysis of subunit interactions. At Bath we are studying the following citrate synthases from several Archaea:

- *Thermoplasma acidophilum*

  Habitat: 55°C, pH 1-3 (internal pH 7.0). We have determined a 2.5 Å citrate synthase structure (Russell *et al.*, 1994) from this organism.
- *Sulfolobus solfataricus*

  Habitat: 85°C, pH 1-2 (internal pH 7.0). Citrate synthase gene cloning in progress (G. Munro).
- *Pyrococcus furiosus*

  Habitat: 100°C, pH 6.5. Citrate synthase gene cloned and sequenced (Muir *et al.*, 1994), and 3.0 Å structure determination in progress.
- *Haloferax volcanii*

  Habitat: 3M NaCl (internal 3M KCl). Citrate synthase gene recently cloned and partially sequenced (G. Munro).

The catalytic parameters of the various citrate synthases are very similar at their normal operating temperatures, i.e. k_{cat} and K_m are about the same for the pig enzyme at 37°C as the *Tp. acidophilum* enzyme at 55°C, and the *P. furiosus* enzyme at 100°C. After incubation at 45°C for 10 min, the pig enzyme is inactive, whereas the *Tp. acidophilum* enzyme remains active after incubation at 80°C for 10 min. CD spectra of the *Tp. acidophilum* enzyme at 25 and 80°C are almost indistinguishable, suggesting preservation of secondary structure.

Information from sequence alone

Analysis of protein thermostability based on sequence information has pointed to certain substitutions on going from mesophile to thermophile (Menendez-Arias & Argos, 1989). These studies, based on thermophilic Bacteria and not Archaea, observed a general increase in hydrophobicity and decrease in flexibility with increasing temperature, with most significance being found in α-helices, both surface exposed and at domain interfaces. They list a 'top-10' substitution list with Lys→Arg and Ser→Ala in the top two places, with a general increase in (Arg)/(Arg+Lys) ratio with increasing temperature. The percentage of each type of amino acid in citrate synthase from pig, *Thermoplasma acidophilum (Tp)* and *Pyrococcus furiosus (P)* are given below:

	C	G	P	A	V	L	I	F	M	W	Y	H	S	T	N	Q	K	R	D	E
Pig	1	7	5	7	6	12	4	3	3	2	4	3	7	5	4	4	6	4	5	5
Tp	0	7	4	11	6	6	8	4	3	1	5	2	5	5	4	3	8	5	5	8
P	0	8	4	7	6	10	9	3	2	2	7	2	5	4	3	1	9	4	3	10

Significant changes are highlighted, and include: (i) a large increase in alanine content from the pig enzyme to the *Tp. acidophilum enzyme* with a concomitant reduction in leucine, a trend which reverses back again for the *P. furiosus* enzyme, (ii) an increase in the number of isoleucines with temperature, (iii) an increase in lysine and glutamic acid content with temperature, (iv) a reduction in thermolabile residues with temperature (Asn, Gln and Cys), and (v) an increase in tyrosines with temperature.

***Thermoplasma acidophilum* citrate synthase**

The following summarises the structure determination of this enzyme:

- 20% sequence identity with pig.
- Crystals of open form: $P2_1$ a=53.8 Å, b=173.8 Å, c=86.7 Å, β=97 °
- 2 dimers of 2 x 384 amino acids / asymmetric unit.
- Solved using AMoRE with a truncated, polyalanine pig dimer model.
- Refined using X-PLOR, R=19% at 2.5 Å

Full details can be found in Russell *et al.* (1994). A comparison with the pig enzyme gave an rmsd of 2.27 Å for 356 Cα atoms (Fig. 1). The structure represents the open form (substrate entry/product release) of the enzyme. The *Tp. acidophilum* enzyme consists of two domains: a large domain with 11 helices, four less than in the pig enzyme, and a small domain with 5 helices, the same number as in the pig enzyme. There is a small 3-stranded antiparallel β-sheet towards the N-terminus.

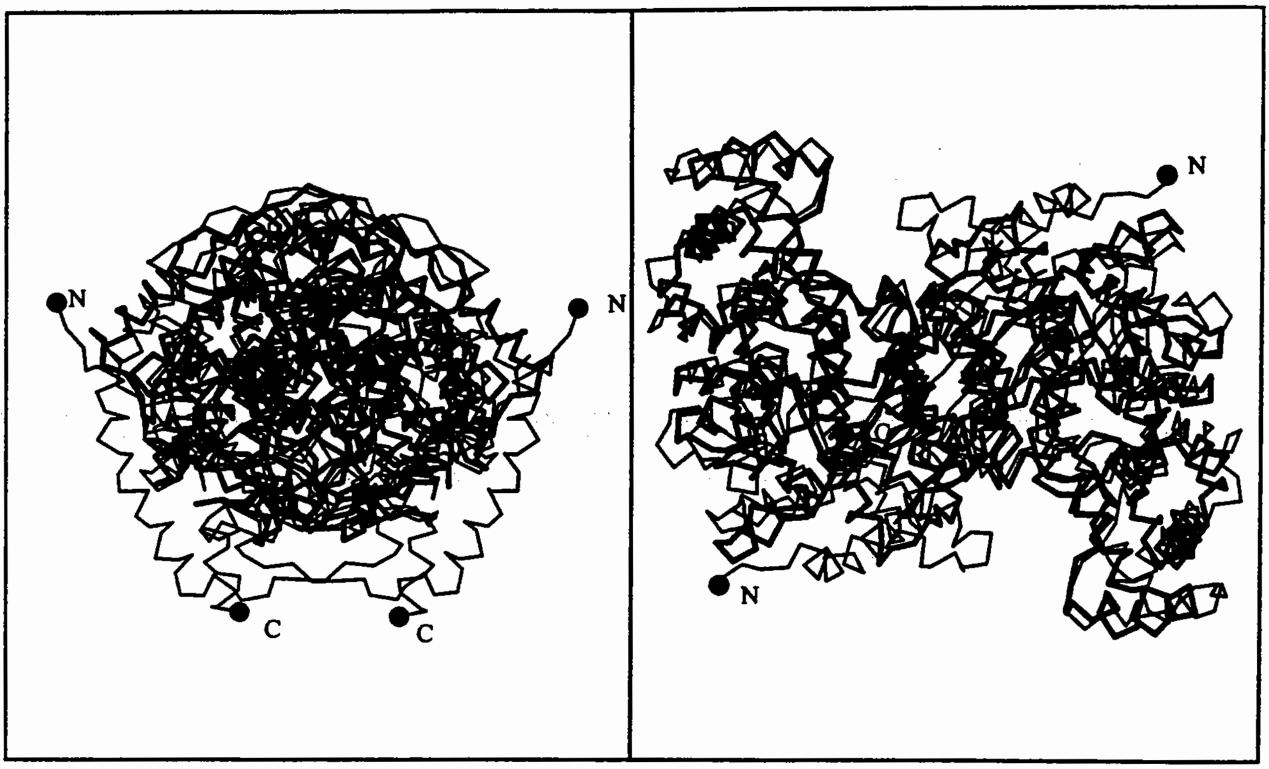
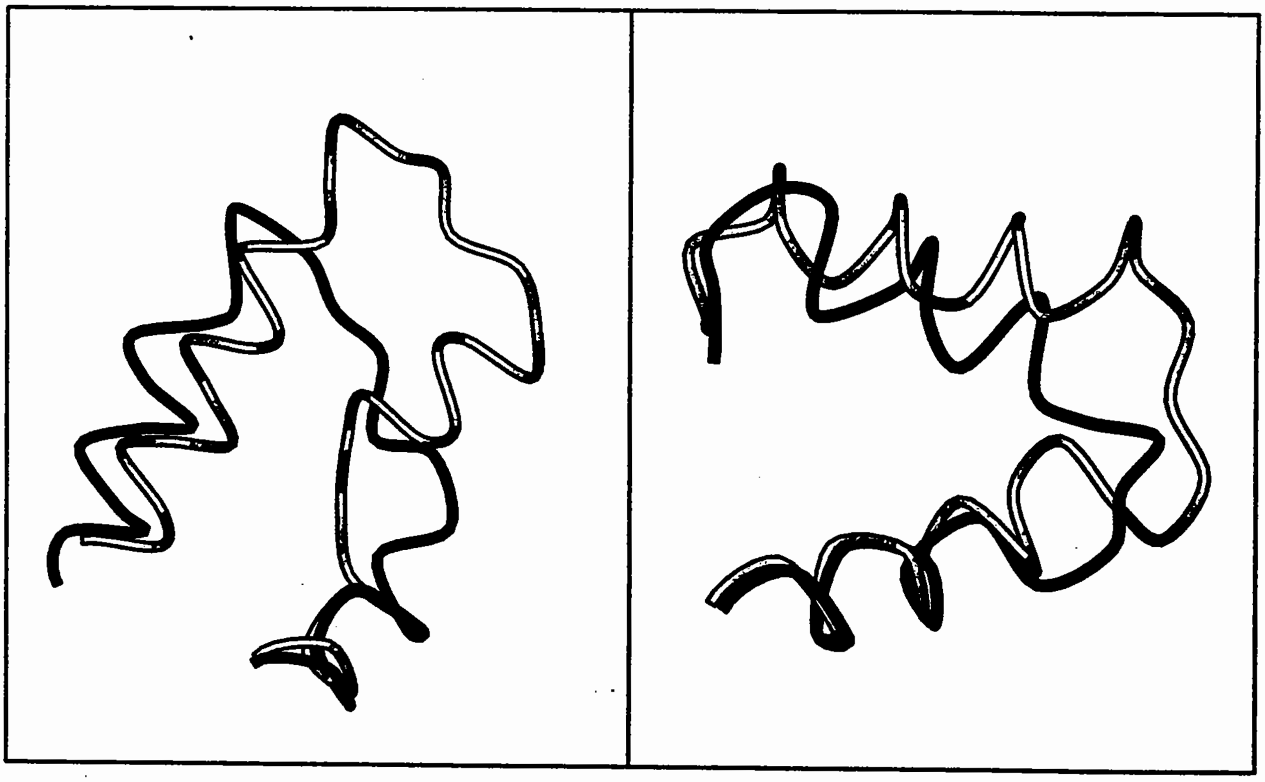


Fig. 1a&1b. Orthogonal views of pig (thin line) and *Tp. acidophilum* (thick line)citrate synthases. The N- and C-termini of the pig monomers are marked.

Fig. 2. Examples of shorter loops in the *Tp. acidophilum* citrate synthase (darker grey).



Possible thermostabilising features of Tp. acidophilum citrate synthase

A comparison of the pig and *Tp. acidophilum* citrate synthase structures gives the following features which might confer thermostability to the Archaeal protein. Alternatively, as the Archaeal protein represents a primitive antecedent of the pig protein, we can think of features which have evolved in the pig protein to give it greater flexibility at mesophilic temperatures, in order to achieve the same catalytic efficiency as that shown by the thermophilic enzyme at its higher operating temperature.

1. N-terminal deletion: The Archaeal citrate synthases are all shorter by around 35 residues when compared to the pig enzyme. This feature cannot necessarily be linked to the enzyme's stability as other non-Archaeal citrate synthases have recently been found to lack the 35-residue N-terminus (Patton *et al.*, 1993; Anderson *et al.*, 1993). Fig. 1a shows the surface location of this 35-residue N-terminus in the pig enzyme where it forms an extensive helix. Its removal, together with the removal of an exposed C-terminal helix in the pig enzyme, leads to a much more compact structure. The role of the N-terminus in the pig enzyme remains unknown.

2. Loop regions: There are four loops that have become truncated in the *Tp. acidophilum* enzyme compared to the pig. Examples are shown in Fig. 2, and one includes removal of helix H from the pig enzyme. It is interesting to note that there are three regions in the open-form of the pig enzyme which have above average temperature factors: the N-terminus and two of the loops that are absent in the *Tp. acidophilum* enzyme. Inclusion of these extra loops into the pig enzyme probably confers some flexibility to the mesophilic protein allowing it either to achieve optimum catalytic activity and/or to provide sites for initiation of denaturation during protein turnover in the cell. Dynamics studies on other proteins (e.g. Daggett & Levitt, 1993) suggest that loop regions undergo the largest deviations, and at higher temperature are likely to be regions of the protein that unfold first during thermal denaturation.

3. Cavities: An analysis of cavities was carried out using VOIDOO (Kleywegt & Jones, 1994), which revealed a marked decrease in the number and total volume of cavities within the thermophilic enzyme compared to the pig. The total volume which can be occupied by a 1.4Å probe was 612$Å^3$ (7 cavities) for the pig dimer, and 211$Å^3$ (4 cavities) for the *Tp. acidophilum* dimer. Fig. 3 shows two such cavities: in both cases a tyrosine is substituted for an alanine or a serine to fill partially the void. It is well established from mutagenesis experiments that cavity creating mutations decrease protein stability (Erikkson *et al.*, 1992), and in a few cases it has been shown that cavity-filling mutations can increase protein stability (Ishikawa *et al.*, 1993).

4. Subunit interface: The dimer interface is made up of four pairs of antiparallel helices FF',GG',LL' and MM'. In the pig and *Tp. acidophilum* citrate synthases, helices F,L and M are hydrophobic on their interface surfaces. However helix G, which is polar in pig enzyme, becomes markedly hydrophobic in the *Tp. acidophilum* enzyme by substitution of serine and threonine by alanine. This reflects earlier findings that alanines are the most helix stabilising residues, and that such substitutions occur at domain interfaces (Menendez-Arias & Argos, 1989). Fig. 4 shows helical wheels for helix G.

5. Aromatics: There is a marked increase in the number of aromatic residues present in the thermophile enzyme, particularly in the small domain. Such features have been noted in other thermostable proteins (Burley & Petsko, 1985).

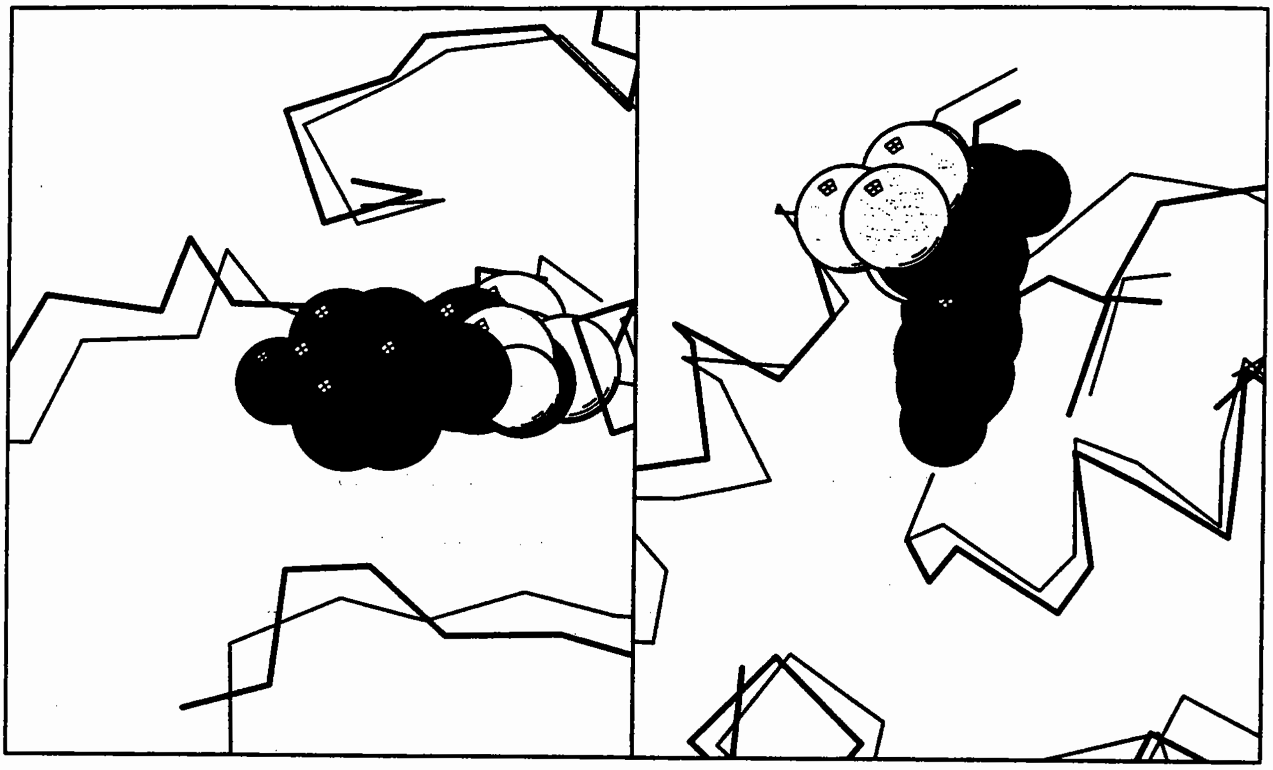
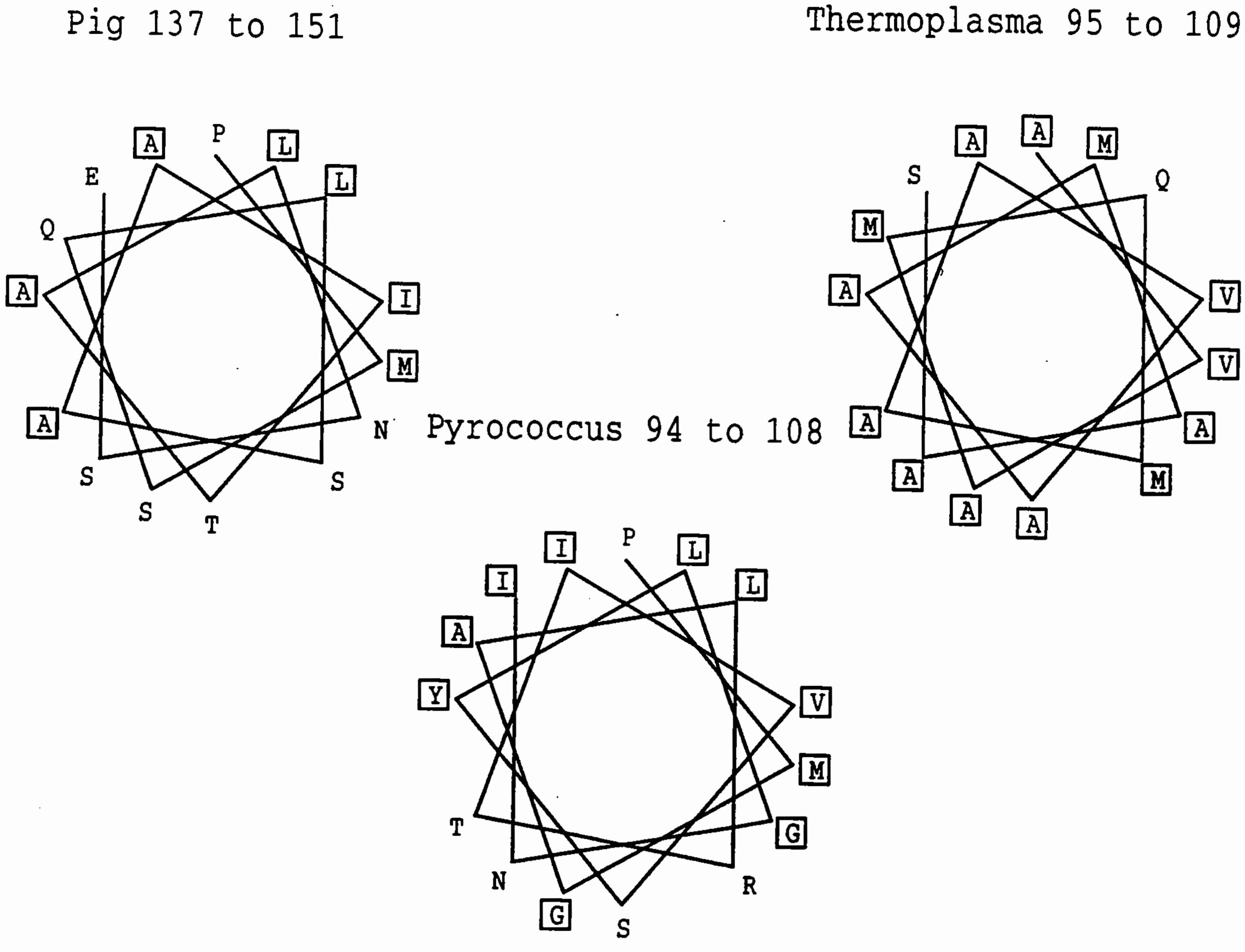


Fig. 3. Two examples of cavity filling substitutions. On the left a serine in the pig enzyme (thin line/light grey) is replaced by a tyrosine in the *Tp. acidophilum* enzyme(thick line/dark grey). On the right an alanine is again replaced by a tyrosine in the *Tp. acidophilum* enzyme.

Fig. 4. Helical wheels for helix G in citrate synthase.



Pyrococcus furiosus citrate synthase

The following summarises the structure determination of this enzyme from an extreme thermophile:

- Gene cloned, sequenced and expressed (J. Muir *et al.*, 1994)
- 42% identity with *Tp.acidophilum*, 31% with pig citrate synthase.
- Crystals in 22 of the 'magic 50'
- $P4_12_12$ a=100.2 Å, c=246.2 Å
- One dimer of 2 x 376 amino acids / asymmetric unit.
- Solved by AMoRE using a combined model consisting of a *Tp. acidophilum* citrate synthase dimer and a polyalanine, truncated pig citrate synthase dimer.

At the time of writing, the structure interpretation is almost complete at 3.5 Å. As the structure has not been refined at higher resolution, a detailed analysis as given above for the *Tp. acidophilum* enzyme will not be given. However, the 42% sequence identity allowed construction of an homology model, based on the *Tp. acidophilum* enzyme structure. This tentatively highlights certain features:

1. Loop regions: The pig enzyme contains 437 residues, *Tp. acidophilum* 384 residues and *P. furiosus* 376 residues. Both Archaeal enzymes are missing the 35 N-terminus residues, and the hyperthermophile is 8 residues shorter than the moderate thermophile. The *P. furiosus* citrate synthase model suggests a further shortening of one or two loops.

2. Subunit interface: Helix G in *P. furiosus* citrate synthase becomes polar, as in the pig enzyme. The dramatic increase in hydrophobicity in this helix in *Tp. acidophilum* citrate synthase is lost. *P. furiosus* must have evolved other mechanisms for maintaining dimer stability at 100°C. *P. furiosus* citrate synthase has a dramatic increase in glutamate residues. The model predicts that three glutamic acid residues on helix F at the subunit interface face their symmetry equivalents. This apparent instability can be understood when we consider the cellular milieu of *Pyrococcus* which contains up to 0.6M KCl together with various 'compatible organic solutes' (Scholz *et al.*, 1992). Indeed, optimal enzyme activity was observed for *P. furiosus* citrate synthase with a buffer containing 0.2M KCl. The crystallisation buffer also contained 0.2M KCl. Its is possible that the potassium ions form strong salt-bridges with the glutamic acid residues thereby stabilising the dimer.

The increased number of glutamic acid residues appear to cluster on the surface of the model, a possible feature observed in halophilic Archaeal proteins. This suggests a 'Centre Pompidou' picture of stabilisation with salt-bridges forming an external series of braces on the surface of the protein. Structure elucidation of the halophilic citrate synthase should shed more light on this theory.

Tp. acidophilum glucose dehydrogenase

We have also determined the crystal structure of glucose dehydrogenase (GDH) from *Tp. acidophilum*. This represents the first glucose dehydrogenase structure from any organism. A summary of the crystallographic analysis is given below; for full details see John *et al.* (1994):

- Crystals: $P2_1$ a=81.8Å, b=120.1Å, c=88.4Å, β=104 °
- 222 tetramer (4x352amino acids) / asymmetric unit.
- Solved by SIR and averaging (RAVE).
- Refined using X-PLOR: R=19% at 2.9 Å.

The structure was found to have a structural similarity to the dimeric liver alcohol dehydrogenase (LADH), which was unexpected from sequence alone. This similarity was observed early on in the structure determination, and allowed an improved envelope description for the averaging. GDH can be thought of as a dimer of LADH-like dimers. The LADH dimer is formed by the two nucleotide-binding domains abutting via their 6-stranded β-sheets to form a twisted 12-stranded sheet.

As there is no mesophilic homologue with which to compare the *Tp. acidophilum* GDH, a detailed analysis of the factors which may confer thermostability cannot be made. However, GDH does contain a classical nucleotide-binding domain, and an analysis of this compared to the many examples in the databank reveals that GDH possesses the smallest such domain. The GDH nucleotide-binding domain contains 114 residues, the same domain in LADH contains 126. This domain in GDH also has a significant increase in the number of aromatic residues: 12 in GDH to 8 in LADH, and appears to contain the highest proportion of aromatics among all nucleotide-binding domains.

The subunit interface in the dimer follows a similar trend to that seen in the *Tp. acidophilum* citrate synthase. In LADH there are 9 intersubunit hydrogen bonds; in GDH there are only four, but in GDH there is a marked increase in the hydrophobic nature of the interface which includes four aromatic residues. Additional potential thermostabilising features include a possible increase in the number of helix-capping residues, and a decrease in the number of thermolabile residues. In LADH, and most members of the Zn-ADH family, there are 6 cysteines involved in co-ordinating two zinc ions per monomer: 2 to the catalytic zinc and 4 to the structural zinc. GDH contains 1 and 3 respectively.

Summary

Analysis of the crystal structures of proteins for a temperature series from mesophile to hyperthermophile is revealing factors that may confer thermostability. Our analysis of citrate synthase, has suggested that thermophilic proteins may be more compact and better packed within their cores. In addition, moderate thermophiles (around 55°C) may utilise stronger hydrophobic interactions for stability, whereas the hyperthermophiles (up to 110°C) may employ external salt-bridges and counterions to provide stability. This may reflect the relative contributions of enthalpy and entropy to protein folding with temperature: at mesophilic temperatures, folding is entropy driven; at high temperatures, the entropic contribution is reduced and folding is enthalpy driven (Dill, 1990). We must await the determination of other temperature families of proteins, and the outcome of our mutagenesis experiments, to see how general our observations are, and whether any rules can be derived for engineering thermostability into biotechnologically useful proteins.

Acknowledgements

We are grateful to the BBSRC for support in the form of studentships to RR, JR and J. Muir, and a Biotechnology Directorate Grant, to MD, DH and GT, now supporting RR and G. Munro.

References

Anderson, S.C.K., Powrie, R. & Mitchell, C.G. Biochem. Soc. Trans. **22** (1993), 415.
Burley, S.K. & Petsko, G.A. Science **229** (1985), 23--28.
Daggett, V. & Levitt, M. J. Mol. Biol. **232** (1993), 600-619.
DeLong, E.F., Wu, K.Y., Prezelin, B.B & Jovine, R.V.M. Nature **371** (1994), 695-697.
Eriksson, A.E., et al., & Matthews, B.W. Science **255** (1992), 178-183.
Horovitz, A., Matthews, J.M. & Fersht, A.R. J. Mol. Biol. **227** (1992), 560-568.
Ishikawa, K., Nakamura, H., Morikawa, K. & Kanaya, S. Biochemistry **32** (1993), 6171-6178.
John, J., Crennell, S.J., Hough, D.W., Danson, M.J. & Taylor, G.L. Structure **2** (1994),385-393.
Kleywegt, G.J. & Jones, A.T. Acta Crystallogr. **D 50** (1994), 178-185.
Menendez-Arias, L. & Argos, P. J. Mol. Biol. **206** (1989), 397-406.
Muir, J.M., Hough, D.W. & Danson, M.J. Sys. Appl. Microbiol. **16** (1994), 528-533.
Patton, A., Hough, D.W., Towner, P. & Danson, M.J. Eur. J. Biochem **214** (1993), 75-81.
Remington, S.J., Wiegand, G. & Huber, R. J. Mol. Biol. **158** (1992), 111-152.
Russell, R.J.M., Hough, D.W., Danson, M.J. & Taylor, G.L. Structure, **2** (1994) 1157-1167.
Woese, C.R., Kandler, O. & Wheelis, M.L.. Proc. Natl. Acad. Sci. USA **87**, (1990) 4576-4579.

Adenylate kinase, structures to mechanism

Georg E. Schulz
Institut für Organische Chemie und Biochemie,
Albertstr. 21, 79104 Freiburg im Breisgau, Germany

The analysis of adenylate kinases is one of my long term interests. The first structure of an enzyme of this family was elucidated as many as 20 years ago (Schulz et al., 1974). At that time we would have just wondered about the present conference slogan "how to make most of your model"; because we were busy enough to build it in wire or plastic and to put it on display. Any further analyses tended to be on the chemical side. Now, with the arrival of gene technology and the publication of numerous protein structures, we take a new structure as a starting point and proceed to further structural analyses.

One of these is the determination of internal forces of the protein as far as they can be deduced. In this respect, I like to present a hydrogen bond analysis in α-helices as an example (Abele, 1994). It is known for some time that hydrogen bonds in β-sheets are significantly shorter than in α-helices and should therefore be stronger. A closer look, however, shows that this deduction may not be true at all, because all α-helical hydrogen bond donors are also involved in secondary hydrogen bonds with neighboring acceptors in addition to those to the primary acceptor. Consequently, "α-bonds" are probably stronger than the shorter but isolated "β-bonds".

Another analysis of internal forces can be pursued by mutating a protein and determining thermodynamic parameters together with the respective structures. We have done this with an adenylate kinase and found out, that the structural changes on mutation are usually rather local and minor, such that even an accurate structure analysis stands not much of a chance in explaining the stability changes. This type of analysis may not be the most rewarding field for a structure lab.

Somewhat more rewarding is the structural analysis of multiple mutations executed in protein engineering endeavors. Here, we took part in the analysis of a cofactor specifity change from NADP to NAD in glutathione reductase (Scrutton et al., 1990; Mittl et al., 1994). Although this change involved seven point mutations, the structural differences remained rather local and small. At the end, the smallishness of the changes prevented an eye-catching demonstration of the superiority of a structure analysis versus were model building.

As a very particular side-result structure analyses yield the crystal packing contacts. These are generally considered as a nuisance, because they perturb the molecule locally. And exactly for this reason we have reported these contacts in all our analyses over the years (Thieme et al., 1981). The number of other labs joining this effort in their publications has remained small. One reason of this limited interest may have been the complexity of a thorough crystal contact analysis, which becomes mind-boggling for more than two molecules in the asymmetric unit. Meanwhile, I view these contacts differently: They provide us with examples for protein associations, which become of increasing importance for the analysis of molecular interactions in the cell. Many of these interactions are rather weak, like the protein interactions in packing contacts. Actually, with the enzyme NADH peroxidase we encountered an example of a crystal contact that is probably also an intracellular contact (Stehle et al., 1991).

A further example for "making more of your model" are structure analyses of homologous proteins under different chemical conditions. In the case of thioredoxin reductase, for instance, the present structural knowledge points to an extreme displacement of the NADP-binding domain with respect to the FAD-binding domain (Kuriyan et al., 1991). Any prove for this displacement, however, requires further structure analyses. In the adenylate kinases the first structure analyses failed to show the catalytic center at any detail. Only the general position was clear from the deep cleft in the protein and the multitude of positively charged residues that could accomodate the four phosphates of the substrates; the enzyme processes ATP and AMP to ADP and ADP with the help of a divalent cation (usually Mg^{2+}, forward reaction). Now after many more structures, we know that the active center is not static; it is dis- and resassembled during each catalytic cycle.

In order to improve our knowledge on the enzyme we analyzed further species with various ligands. This endeavor was limited by the success of protein crystallization, which was not overwhelming at all. A decisive advance was then made by using the inhibitor Ap_5A, which is ATP and AMP connected by a fifth phosphate. With further structures coming up (Müller & Schulz, 1992), we were surprised about the large conformational differences between them and suggested that each catalytic cycle involves large motions (Schulz et al., 1990). Until now we succeeded with the structure analysis of about two dozen crystals of adenylate kinases with various ligands.

After comparing several adenylate kinases it became clear that the chain fold should be subdivided into three domains: (i), the CORE domain of about 140 residues containing the

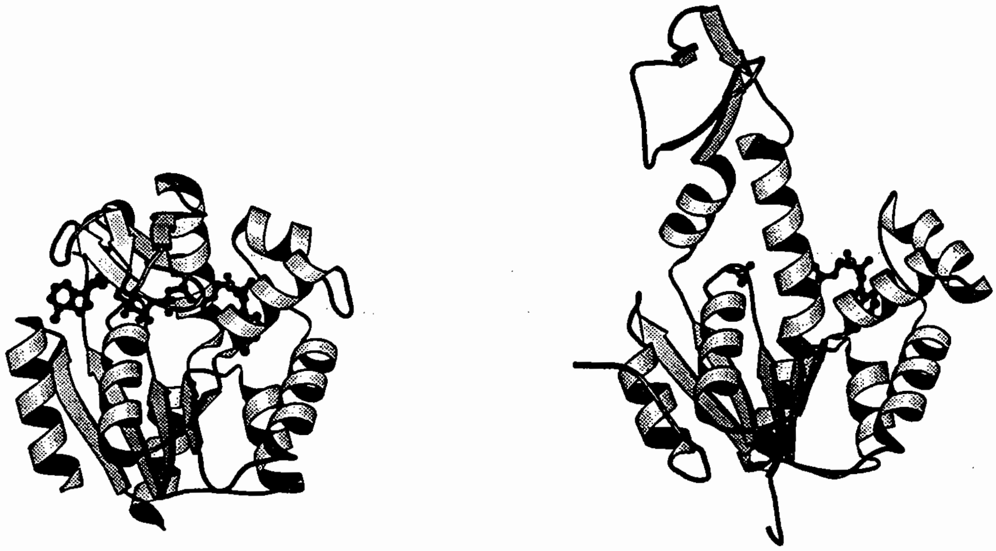


Figure 1. Ribbon plots of the adenylate kinase from E.coli with bound substrates ATP and AMP (in fact the two-substrate-mimicking inhibitor Ap$_5$A) and of the bovine mitochondrial matrix adenylate kinase with bound AMP. A comparison reveals the motion on ATP binding.

central parallel β-sheet with surrounding α-helices; (**ii**), the NMPbind domain of about 30 residues, which can be considered as being inserted in the chain after about 35 residues; and (**iii**), the LID domain inserted after about 100 residues. From an architectural point of view CORE is the stable base, NMPbind encloses the bound AMP (in general NMP) and LID covers the bound ATP. From structures with ligated AMP alone, ligated ATP-analogue alone, and both ligated substrates (the analogue Ap$_5$A) we found that domains NMPbind and LID move rather independently relative to CORE on binding the respective ligand. From steady state kinetics it is known that there exists no compulsory binding sequence of the nucleotides before catalysis.

A more detailed analysis of the NMPbind motion showed that it carries the backbone over distances of up to about 15 Å. It is not a pure rotation, but can be reasonably described by a rotation coupled with a shearing motion. AMP is completely encapsulated, we observe here an immense induced-fit on binding this substrate.

One of our homologous structures is a uridylate kinase ligated with AMP instead of UMP (Müller-Dieckmann & Schulz, 1994). This enzyme phosphorylates AMP and CMP as side-activities. Consequently, it has assumed multiple NMP specifity in spite of the large induced-fit. A closer look at the structures and the amino acid exchanges between adenylate and uridylate kinases at this site showed that UMP and CMP most likely bind in combination with a water molecule and thus need a somewhat larger pocket than AMP. This additional water molecule cannot be squeezed out by the induced-fit motion, because in the uridylate kinases (and only in this species) the induced-fit motion is blocked as domain NMPbind runs into domain CORE at a pocket size suitable for UMP plus the water molecule. Accordingly, the structural analyses have shown how a highly specific AMP site undergoing a large induced-fit motion can be reorganized to multiple specificity in spite of this motion.

The LID motion is easier to describe than the NMPbind displacement because it is mostly a rotation of a 40-residues domain that moves as a solid block carrying the backbone over distances of up to 35 Å. After this movement, LID covers the bound ATP rather well but not completely. Accordingly, the notion "induced-fit" is not quite appropriate in this case. Moreover, there are a number of socalled short variants of the NMP-kinases containing a LID of only 10 residues that provides not much of a cover for ATP at all.

The LID motion has been analyzed by superimposing parts of the structures for finding the hot spots of change (Gerstein et al., 1993). This showed that there are four of such hot spots, two at each of the two chain connections between LID and CORE. These connections are α-helices and the hot spots are before and after them. Three of the four hot spots are essentially rotations around single bonds, whereas the fourth one involves rotations around several bonds and contains two strictly conserved aspartates. The fourth hot spot is also present in the short variants with their minute LID.

The conserved aspartates undergo large conformational changes on ATP binding as they turn around to fix arginines that attach to the phosphates. Accordingly, the incoming ATP triggers the multiple-torsion-angle changes at the major hot spot, which most likely cause the motion of LID. The other three hot spots are probably passive elements as they involve merely single torsions. Moreover, this rearrangement is required for catalysis, because the fixed arginines are. We therefore conclude that ATP triggers the assembly of the catalytic center, which is in a disassembled state if no substrates are present. These observed motions could be a paradigm for actions during energy transductions in muscle and mitochondria, because the very peculiar way of binding ATP with its β-phosphate in a giant anion hole formed by the polypeptide backbone (Dreusicke & Schulz, 1986) is also found in myosin (chemical to mechanical) and in F_1-ATPase (osmotic & electrochemical to chemical).

The structure comparisons provided also information on the catalytic mechanism. It is known from isotope labeling experiments (Richard & Frey, 1978) that the NMP-kinases use an in-line S_N2 mechanism for the transfer of the phosphoryl group since there occurs an inversion of its chirality. Such a transfer could be associative or dissociative; the latter requires a metaphosphate as a transition state. Superpositions of our structures using the CORE domains brought ligated ATP and AMP always within 1 Å at the same position, which demonstrated that all analyzed members of the NMP-kinase family use the same mechanism. It showed further that the α-and β-phosphate of ATP and the α-phosphate of AMP were particularly well-conserved at their positions. We therefore proceeded by superimposing the structures merely on these phosphates in order to work out the residual differences. This superposition showed that the phosphate accepting/donating oxygen atoms are at a distance of 4.7 Å, which is a 1.6 Å O-P bond length plus a 3.1 Å O-P van der Waals distance and therefore just right for an associative transfer, excluding the dissociative alternative. Moreover, these two oxygens are placed in ideal geometry for the transfer.

Given this geometric detail, the transition state can be circumscribed by a trigonal bipyramid suspended between these two anchoring oxygens. Since this state is unstable, of course none of our stable structures can assume it. With so many ligated structures available, however, we reach a situation where this elusive transition state can be approximated by an average of all observed structures under the assumption that all perturbations are random and therefore average out. Indeed, the average of all γ-phosphate orientations of our structures yield an O-P bond for the γ-phosphate of ATP that deviates by only 7° from the central line of the bipyramid. Thus, averaging appears to be a viable concept of establishing the fine details of a transition state.

Another clear result of the phosphate superposition is the observation that in all structures with ligated Ap_5A always the fourth phosphate shows the largest perturbation, whereas the third one (γ-phosphate of ATP) is much less displaced. Therefore, all our structures are closer to the start of the forward than to the start of the backward reaction. This observation is corroborated by a structure that contains the magnesium ion required for catalysis. This Mg^{2+} is most likely at its correct place, because a phosphate superposition with GTP-analogues bound to G-proteins causes all Mg^{2+} ions to coincide. Again, this is the place of Mg^{2+} for the forward reaction.

In the G-proteins Mg^{2+} has a very regular octahedral ligand sphere with two water, two peptide and two phosphate ligands. In contrast, Mg^{2+} in the NMP-kinases has a distorted octahedral sphere with three water and two phosphate ligands. The sixth ligand is absent, leaving space for the Mg^{2+} to accompany the phosphoryl group during transfer. This is all set for the forward reaction. For the backward reaction, however, a similar octahedron can only be found after a water rearrangement. Apparently all our structures favor the start of the forward reaction, although the two reaction rates are similar as the equilibrium constant is about one.

With so many structures available that contain different ligands and form various crystal contacts, we have a broad collection of intermediate states of the movements of both, the NMPbind and the LID domains. These intermediates are exactly what one would need to produce a movie recording these motions. We therefore assumed that the observed structures are indeed still pictures of a movie. We sorted the structures according to the progress of the movements and displayed them one after the other, creating the movie of a catalytic cycle. In order to make the motions smoother, we eliminated some of the structures that were far out, possibly because of crystal packing effects, and we interpolated between some of the real structures producing approximately equal-spaced intermediates.

The resulting movie records the motions during one catalytic cycle which is about one millisecond as derived from the turnover number of this enzyme family. A special attraction of the adenylate kinases are the very large motions in comparison to the size of the molecule. Thus, we achieved a very dynamic representation of life by establishing a large number of static pictures. I am sure that there will be many more such movies in the future.

References

Abele,U., Dissertation (1994), Albert-Ludwigs-Universität, Freiburg im Breisgau

Dreusicke,D. and Schulz,G.E., FEBS-Letters *208*(1986)301

Gerstein,M., Schulz,G.E. and Chothia,C. J.Mol.Biol. *229*(1993)494

Kuriyan,J., Krishna,T.S.R.; Wong,L., Guenther,B., Pähler,A., Williams,C.H.Jr. and Model,P., Nature *352*(1991)172

Mittl,P.R.E., Berry,A., Scrutton,N.S. and Perham, R.N.
Protein Science *3*(1994)150

Müller,C.W. and Schulz,G.E., J.Mol.Biol. *224*(1992)159

Müller-Dieckmann,H-J. and Schulz,G.E., J.Mol.Biol. *236*(1994)361

Richard,J.P. and Frey,P.A., J.Amer.Chem.Soc. *100*(1978)7757

Schulz,G.E., Elzinga,M., Marx,F. and Schirmer,R.H.
Nature *250*(1974)120

Schulz,G.E., Müller,C.W. and Diederichs,K.
J.Mol.Biol. *213*(1990)627

Scrutton,N.S., Berry,A. and Perham,R.N., Nature *343*(1990)38

Stehle,T., Ahmed,S.A., Claiborne,A. and Schulz,G.E.
J.Mol.Biol. *221*(1991)1325

Thieme,R., Pai,E.F., Schirmer,R.H and Schulz,G.E.
J.Mol.Biol. *152*(1981)763

CONCEPTS, DIFFICULTIES AND PROGRESS IN STRUCTURE BASED DRUG DESIGN

E. P. Mitchell[1], K. A. Watson[1], C. Bichard[2,3], G. W. J. Fleet[2,3], S. E. Zographos[4], N. G. Oikonomakos[4], M. Board[5] & L. N. Johnson[1,2]

[1]Laboratory of Molecular Biophysics, University of Oxford, Oxford, OX1 3QU, UK and [2]Oxford Centre for Molecular Sciences, Oxford; [3]Dyson Perrins Laboratory, Oxford; [4]National Hellenic Research Foundation, Athens; [5]Biochemistry Department, Oxford.

Summary

Protein crystal structure determinations have made important contributions to understanding the molecular basis of recognition and function for molecules important in medicine and have led, in some selected examples, to the design of new therapeutic agents. This paper summarises recent achievements and reports on progress in our long term study to provide potential agents for the treatment of diabetes. Structural studies with glycogen phosphorylase have led to the design of a series of glucose analogue inhibitors that can act as powerful regulators of liver glycogen metabolism. The best compound to date has involved some novel carbohydrate chemistry and exhibits a K_i that is 3 orders of magnitude lower than that of glucose. Physiological studies have shown that inhibitors more potent than glucose produce dramatic effects in inhibition of glycogen degradation and promotion of increased glucose utilisation and glycogen deposits in isolated hepatocytes.

Introduction

The earliest known drugs were those derived from natural products whose beneficial effects were detected from observation. Thus aspirin, from willow bark, was used as an analgesic long before its pharmacological properties or molecular basis of action was understood. The conceptual breakthrough, developed in the nineteenth century, that a drug may have specific target at the cellular level led to the notion of receptor based drug design that has profoundly benefited the drug discovery process. The strategy in the design of new agents has been to modify the structure of a lead compound by systematic chemistry in conjunction with biological and physiological evaluation to produce a compound of the required potency. The most recent phase in drug design has utilised knowledge of the three dimensional structures of target molecules or of related compounds. From experimental observations at the atomic level of how inhibitors bind to their macromolecules, specific interactions that are important in molecular recognition can be inferred. This knowledge can be applied to lead compounds in *de novo* design as well as to improvement in existing leads and can provide insights into mechanisms of existing drugs. However although much attention has been focused on the potential of structure based drug design the number of compounds that have neared the market place is limited.

The design of tight inhibitors has been impressively accomplished but the more stringent requirements that a drug must fulfil in order to treat patients such as bioavailability, toxicity, metabolism, half life and cost effectiveness are more difficult to address. Structure based drug design has a long way to go before it can rival the products such as cortisones, anti-inflammatory agents, antibiotics, antidepressants and hypoglycaemic agents that were produced by more conventional methods. But it provides an alternative approach. It has been estimated(Vagelos, 1991) that on average in a trial and error procedure some 10000 compounds will be screened, 10 will go forward to trials and 1 may become a prescription medicine. That 7 out of 10 medicines do not recoup their expenses and that fewer than 5 earn more than $1B per year. There are diseases of the rich and diseases of the poor and the majority of life threatening diseases occur in poor countries where the likelihood of regaining costs is small. Many diseases are a product of poor social and health conditions. The development of cures lies not only in providing protection and therapy but also in improving living standards. It is to be hoped that the social cures may be forthcoming and that in addition the new knowledge created by the crystallography will also contribute.

Current achievements in structure based drug design

There has been a host of structural results on proteins of medical importance (Perutz, 1992). They include studies that contribute to understanding how existing agents work and which may lead to new compounds with improved properties. Examples include work with acetyl cholinesterase for agents that block neurotransmission, with bacterial DNA gyrase for understanding the action of some antibiotics, with HIV reverse transcriptase for understanding the mechanism of AZT and other potential drugs, with thrombin and tissue plasminogen activator for regulation of blood coagulation, and with prostaglandin synthase for understanding the action of aspirin and other anti-inflammatory agents. Structural data have led to clues for improvement in properties of the anti-viral compounds against human rhino virus. The structure of the cyclophilin A-cyclosporin complex is relevent for understanding the immunosupressant activity of the drug cycolopsorin. Knowledge of protein structure has led to new proteins produced by recombinant DNA technology such as a fast acting insulin for treatment of diabetes or "humanised" antibodies for the treatment of leukaemia and arthritis. Results on the human histocompatibility complex and associated bound peptides, of cytokines and a cytokine receptor complex, super antigens, cell surface adhesion molecules, and the outstanding work on the human growth hormone receptor complexed with the hormone and the tumour suppressor gene product p53 complexed with DNA have provided new insights into basic recognition events in a variety of biological responses.

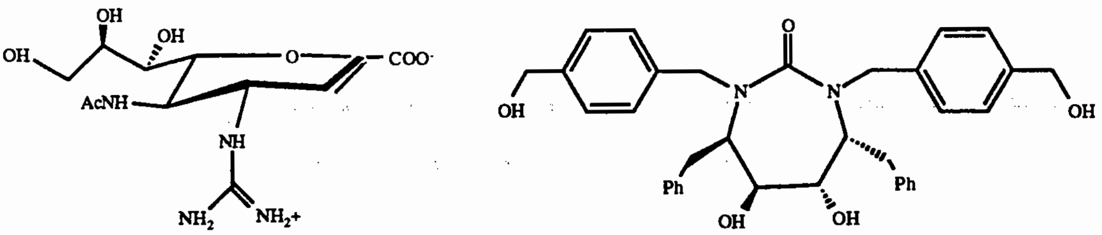
Figure 1 shows 4 examples of compounds that have been designed on the basis of structure. Each of these compounds are in clinical trials. They include:

a) *Influenza virus neuraminidase* inhibitors where knowledge of the structure of the neuraminidase complexed with a transition state compound and the computer programme GRID (Goodford, 1985) were used to design a new compound (Von Itzstein et al., 1993). The addition of a guanidino group to interact with 2 carboxylates resulted in a compound (Fig. 1a) that exhibited a $K_i = 2 \times 10^{-10}$ M , 4 orders of magnitude better than the starting compound and significant *in vivo* activity against influenza virus replication.

b) *HIV protease inhibitors*. In the search for new therapeutic compounds to combat AIDS the HIV protease is probably the most widely studied protein by X-ray methods with over 150 structure determinations of inhibited complexes reported. Recent work has concentrated on non-peptide based inhibitors that have better oral bioavailability. Mechanistic information was used in the design of an agent that exploited the 2-fold symmetry of the dimeric enzyme with a C2-symmetric diol in order to target the 2 aspartates, included the use of a cyclic urea carbonyl oxygen to mimic the hydrogen bond features of a key structural water molecule, and in which selectivity was engineered and restrained by a preorganised scaffold from knowledge of the binding site geometry (Lam et al., 1994). The result was a series of relatively low molecular weight compounds with high oral bioavailability and good potency against the virus. Compound DMP 323 (Fig. 1b) has a $K_i = 2.7 \times 10^{-10}$ M.

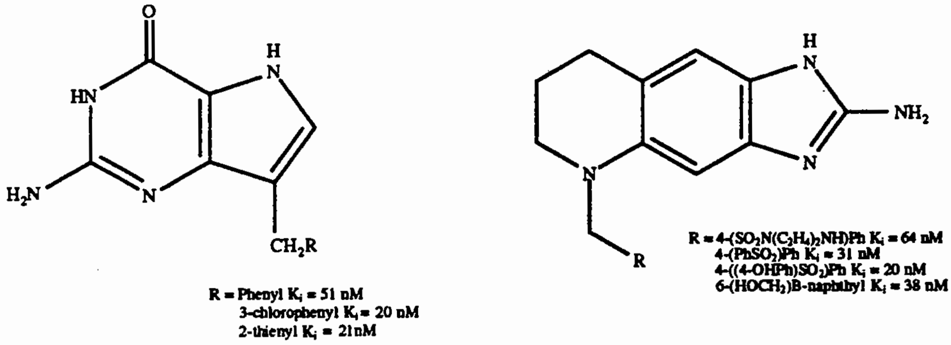
c) *Purine nucleoside phosphorylase:* Interest in this enzyme as a drug target arises from two rather different properties. The enzyme metabolises purine nucleosides, including anticancer agents and anti-AIDS drugs and hence its inhibition might allow these drugs longer action. Secondly inhibitors may have application in the treatment of T-cell proliferative diseases. Starting with the observation that 5'-deoxy-5'-iodo-9-deazainosine was the most potent available inhibitor and that the N-9 position allowed substitution of groups. The design and synthesis of a number of 9-(arylmethyl) derivatives of 9-deazaguanine has been reported (Montgomery et al., 1993)(Fig. 1c). The improved potency of the 9 substituted compounds has been rationalised from X-ray analysis where subtle changes in hydrogen bonding patterns and location of the phenyl group between 2 phenylalanine side chains were observed.

d)*Thymidylate synthase*: Thymidylate synthase is the limiting enzyme in the metabolic pathway for the *de novo* synthesis of thymidylate and as such is a target for drugs against cancer and other proliferative cell diseases. Using the E. coli structure of thymidylate synthase as a model for the human enzyme (70% identity in sequence), a series of 6,7-imidazotetrahydroquinoline inhibitors have been developed (Reich et al., 1992) (Fig 1d). Some of the resulting compounds were shown to effectively inhibit the growth of 3 tumour cell lines *in vitro* .



a) 4-guanidino unsaturated neuraminic acid inhibitor of influenza virus neuraminidase

b) Cyclic urea inhibitor of HIV protease


O
H
N
HN
H_2N
N
CH_2R
R = Phenyl K_i = 51 nM
3-chlorophenyl K_i = 20 nM
2-thienyl K_i = 21nM
H
N
NH_2
N
N
R
R = 4-($SO_2N(C_2H_4)_2NH$)Ph K_i = 64 nM
4-($PhSO_2$)Ph K_i = 31 nM
4-((4-OHPh)SO_2)Ph K_i = 20 nM
6-($HOCH_2$)B-naphthyl K_i = 38 nM

c) 9-(arylmethyl)-deazaguanine inhibitors of purine nucleoside phosphorylase

d) Imidazotetrahydroquinoline inhibitors of thymidylate synthase

Figure 1 Structure based inhibitors of a) influenza virus neuraminidase; b) HIV protease; c) nucleoside phosphorylase and d) thymidylate synthase.

Curiously it appeared that cell growth inhibition may not have been the primary consequence of thymidylate synthase inhibition. As the authors discuss, the information from a protein crystal structure has the potential of increasing selectivity for a given target but having a tight inhibitor does not necessarily result in the specific targeting of that enzyme in cells. Factors which are difficult to predict such as binding to other cellular proteins, metabolism or transport properties can render a potent inhibitor ineffective or possibly more effective in cell culture and *in vivo.*

Glycogen phosphorylase and diabetes

The structure of glycogen phosphorylase has been used to design glucose-analogue inhibitors that have increased potency compared to the parent compound glucose and which may prove beneficial in the regulation of glycogen metabolism in Type II diabetes. Diabetes mellitus is characterised by chronic elevated blood glucose levels. The disorder affects 2% of the population in the Western world and 75% of this total is accounted for by the non-insulin dependent form of the disease (NIDDM or Type II diabetes). NIDDM is managed by diet, exercise, hypoglycaemic drugs, which are based on 3rd generation sulphonylureas that were originally developed in the 1930s as antibiotics, and, if these fail, by insulin therapy. The current drugs are not entirely satisfactory and there is a continued interest in new agents that can control blood glucose levels. Hyperglycaemia in NIDDM patients is a result of diminished insulin release and or insulin resistance that leads to impaired tissue glucose uptake and impaired suppression of the output of glucose from liver, even when blood glucose levels are already high (DeFronzo, 1988).

A simplified scheme for the regulation of glycogen in liver is shown in Figure 2. Glycogen concentrations are regulated by the activities of glycogen phosphorylase (GP) and glycogen synthase (GS). Activation of GP by phosphorylation (GPb to GPa) is achieved by the action of a single enzyme, glycogen phosphorylase kinase, at a single site, Ser14, but the reciprocal inhibition of glycogen synthase (GSI to GSD) through phosphorylation is effected by at least 5 different kinases acting on multiple different serine sites. The reverse reactions of inactivation of phosphorylase and activation of glycogen synthase are achieved *in vivo* largely through the action of a single enzyme, protein phosphatase-1G (PP-1G). It is mainly through this enzyme that the co-ordination of glycogen breakdown and synthesis is achieved in response to glucose in liver (Hers, 1976; Stalmans, 1976) and in response to insulin in muscle (Dent et al., 1990). In liver PP-1G is targeted to glycogen by a glycogen binding subunit and the PP1 catalysed dephosphorylation of glycogen synthase is inhibited allosterically by extremely low concentrations (2-20 nM) of the active form of phosphorylase, GPa (Allemany et al., 1986; Wera et al., 1991). GPa does not inhibit its own dephosphorylation by PP-1G (K_m 2 mM) and hence activation of GS is achieved after a lag period during which hepatic GPa is converted to its inactive form GPb by PP-1G and inhibition by GPa of PP-1G activity against GS is relieved.

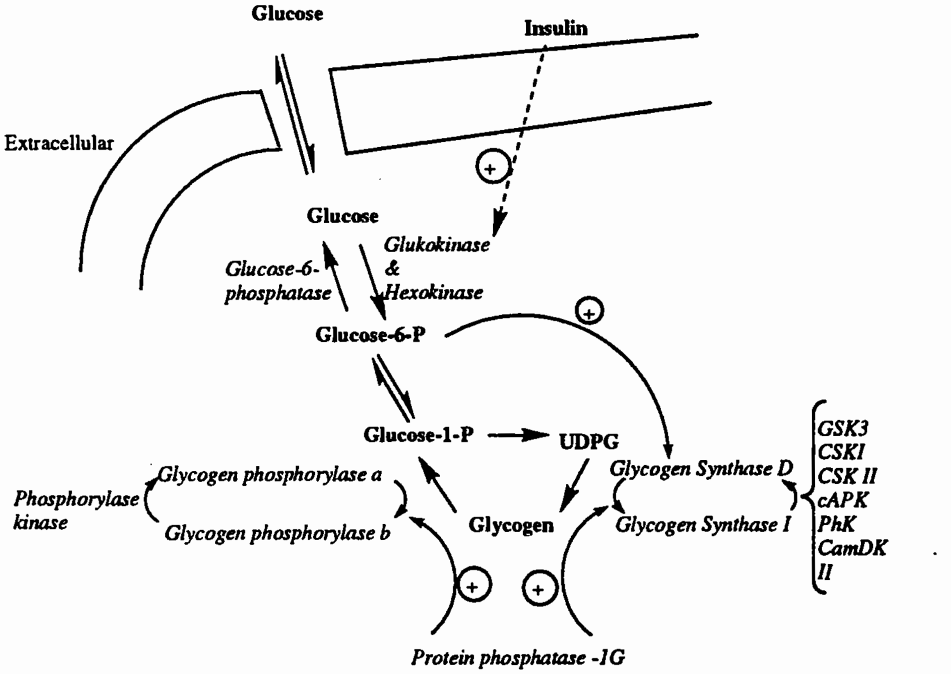


Figure 2. Simplified scheme for regulation of hepatic glycogen metabolism by insulin and glucose.

Glucose is able to augment these effects by competitive inhibition of GPa and by the promotion of the less active T state of GPa which is a better substrate for PP-1G than the active R state GPa. The demonstration that insulin augmented the glucose stimulation of GS and inhibition of GP (Witters et al., 1978) provided an important link between the roles of glucose and insulin in hepatic glycogen metabolism. A rationalisation for the effects of glucose on GP has come from the X-ray crystal structure of the active and inactive forms of the rabbit muscle GP (Sprang et al., 1982; Martin et al., 1990). Glucose is an inhibitor that binds to the catalytic site in competition with substrate but also stabilises the T state (less active) form of the enzyme (in the nomenclature of Monod, Wyman and Changeux) by making specific interactions with a loop of chain (the 280s loop) that blocks access to the catalytic site.

These results have suggested that a more powerful T state inhibitor of GPa than glucose itself, might be of interest in regulation of glycogen metabolism and may provide leads for compounds that could alleviate hyperglycaemia for treatment of Type II diabetes (Martin et al., 1991). A systematic analysis of glucose analogue inhibitors has been carried out using knowledge of the T state structure of rabbit muscle GPb as a model and involving the design and organic synthesis of novel carbohydrate compounds (Martin et al., 1991; Johnson et al., 1994; Watson et al., 1994; Bichard et al., 1995; Watson et al., 1995). A summary of some of these compounds and their K_i values is given in Figure 3 . One of the most effective compounds discovered early on is an N-linked C1 derivative of β–D-glucose (N-acetyl-β–D-glucopyranosylamine (1-GlcNAc)) compound **6** (Figure 3). **6** exhibits a K_i for rabbit muscle GPb of 32 μM, a value which is 200-fold lower than the corresponding K_i of 7 mM for β–D-glucose. In recent work the glucopyranose analogue of hydantocidin, compound **11** (Figure 3), has been found to be an even better inhibitor with a K_i 3 μM, 10^3 times better than the parent compound. This compound was discovered following interest in hydantocidin, a furanose based spirohydantoin. This naturally occurring compound has promise as a herbicide with very little evidence of toxicity to mammals. The corresponding glucopyranose analogue of hydantocidin was modelled into the catalytic site of GPb and found to exploit additional hydrogen bonds to the protein to those made by the parent compound glucose. Accordingly the 2 epimeric spirohydantoins of glucopyranose were synthesised and tested with glycogen phosphorylase to reveal the first specific enzyme inhibition by a spirohydantoin at the anomeric position of the sugar (Bichard et al., 1995). Here we provide a summary of the design and rational for the biochemical activities of the different glucose analogues that led to the synthesis of this potent GP inhibitor.

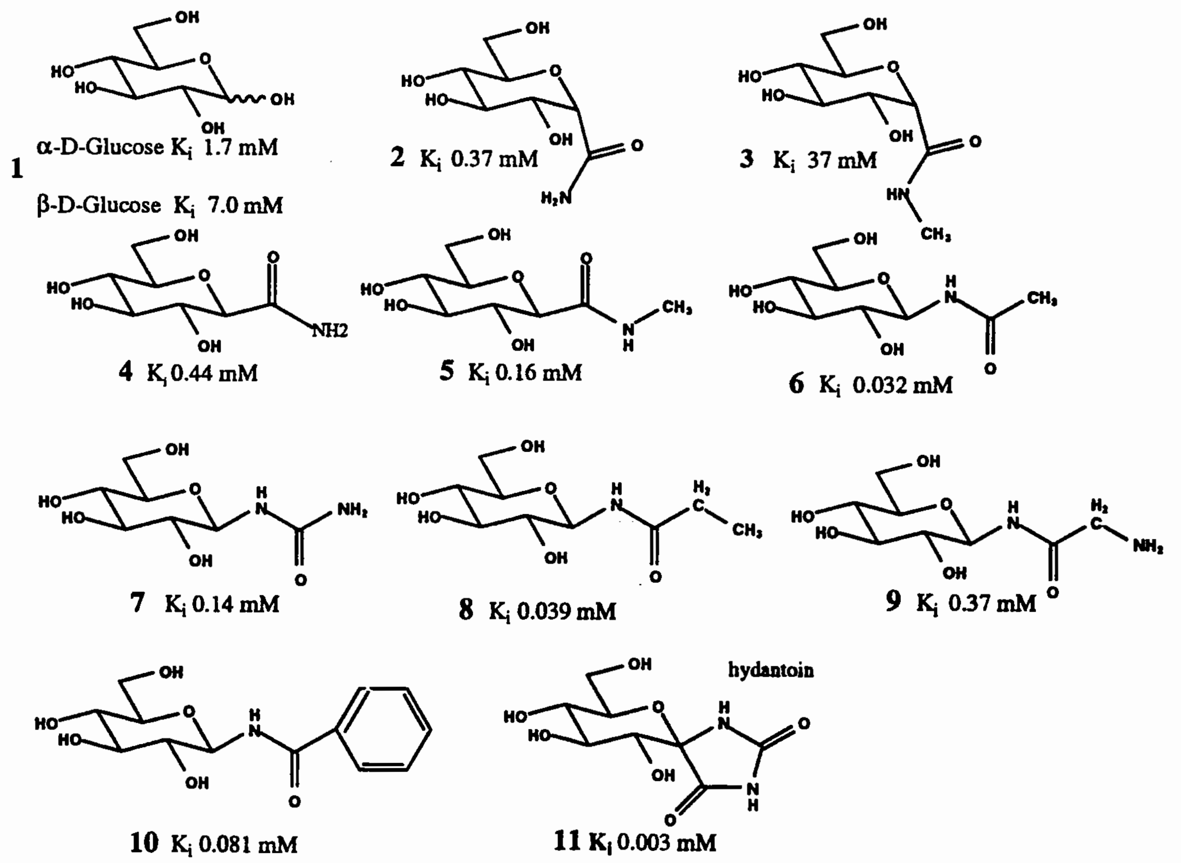


Figure 3. Glucose analogue inhibitors of glycogen phosphorylase and their K_i values.

Crystallographic studies.

In the structural studies, crystal of T state GPb were soaked in solutions containing 100 mM or 50 mM of the compound of interest for 1-2 h and X-ray diffraction data to 2.4 Å resolution collected with typical merging R values of 7-8 %. The structures of the complexes were refined with XPLOR to crystallographic R values of 17-20 % and an estimated error in coordinates of about 0.2 Å, rms deviation from ideal bond lengths of 0.017 Å and angles 3.5°. Examination of the structures with reference to the glucose complex revealed the following observations on the correlation between structure and activity.

(i) *Binding of the lead compound:* α–D-glucose (**1**) binds with each of its peripheral hydroxyl groups involved in hydrogen bonds both as donors and as acceptors and there is little scope for modification at these sites (Figure 4a). There is however a deep pocket adjacent to the C1 atom in the β configuration and a smaller pocket partially blocked by water molecules adjacent to the α configuration. The hydrogen bonds from O2 and α–O1 through water to Asn 284 and Asp 283, respectively, are important for stabilisation of the T state structure (Martin et al., 1990; Martin et al., 1991). Despite these specific polar interactions, glucose binding is relatively weak probably because of few strong van der Waals interactions (there are no interactions with aromatic groups) and the energy cost of transferring a polar molecule from bulk solvent to the catalytic site.

(ii) *Additional hydrogen bonds through water molecules from ligand to protein are important*. The α–heptonamide **2** (K_i 0.37 mM) bound 5 times better than glucose and exploited hydrogen bonds through water molecules from its CO and NH groups to Asp283 (Watson et al., 1994).

(iii) *The conformation of the ligand is important.* Attempts to improve on **2** with an additional methyl group (compound **3**) led to a substantial reduction in affinity (K_i 37 mM). Single crystal studies revealed **3** adopted a skew boat conformation which, when bound to GP, made fewer favourable hydrogen bonds than the corresponding glucopyranose chair. Additional atoms can cause unexpected effects (Watson et al., 1993).

(iv) *Similar K_i values can be generated by quite dissimilar contacts.* Compound **4**, the β–C-amide compound of the β–glucoheptonic acid series had a K_i 0.44 mM similar to **2**, the α–C-amide. But in the complex with **4** the stabilising interactions were not to water but from the amide N to the main chain carbonyl oxygen of His377 and from the CO to Asn 284 side chain (Watson et al., 1994).

(v) *Displacement of water molecules can prove advantageous.* Compound **5**, the methyl analogue of **4**, bound with a K_i 0.12 mM in a similar position to compound **4**. There was no hydrogen bond to Asn 284 but the water molecule, Wat OH4 847, was displaced by the methyl group (Figure 4b). The additional van der Waals contacts to the methyl and the displacement of the water contribute about 0.6 kcal/mol to the binding energy (Watson et al., 1994). The decrease in entropy on transferring a water molecule from liquid water to a macromolecule has been estimated from measurments to be between 0 to 6 kcal mol^{-1} (Bryan, 1987) or 0 to 2 k cal mol^{-1} (Dunitz, 1994) at 20 °C.

(vi) *Linear hydrogen bonds and displacement of more water molecules are even better.* Compound **6**, the β–N-acetyl glucopyranosylamine, (1-GlcNAc), had a K_i 0.032 mM, two orders of magnitude tighter than the corresponding parent β–D-glucose compound and one order of magnitude tighter than the corresponding glucoheptonic acid **5**. The reversal of the amide functionality (**6** vs **5**) led to a shorter, more linear hydrogen bond from the amide nitrogen to the main chain carbonyl oxygen of His 377 (Figure 4c). In addition 2 waters, OH4 847 and OH8 872, were displaced and their displacement appears to be correlated with the

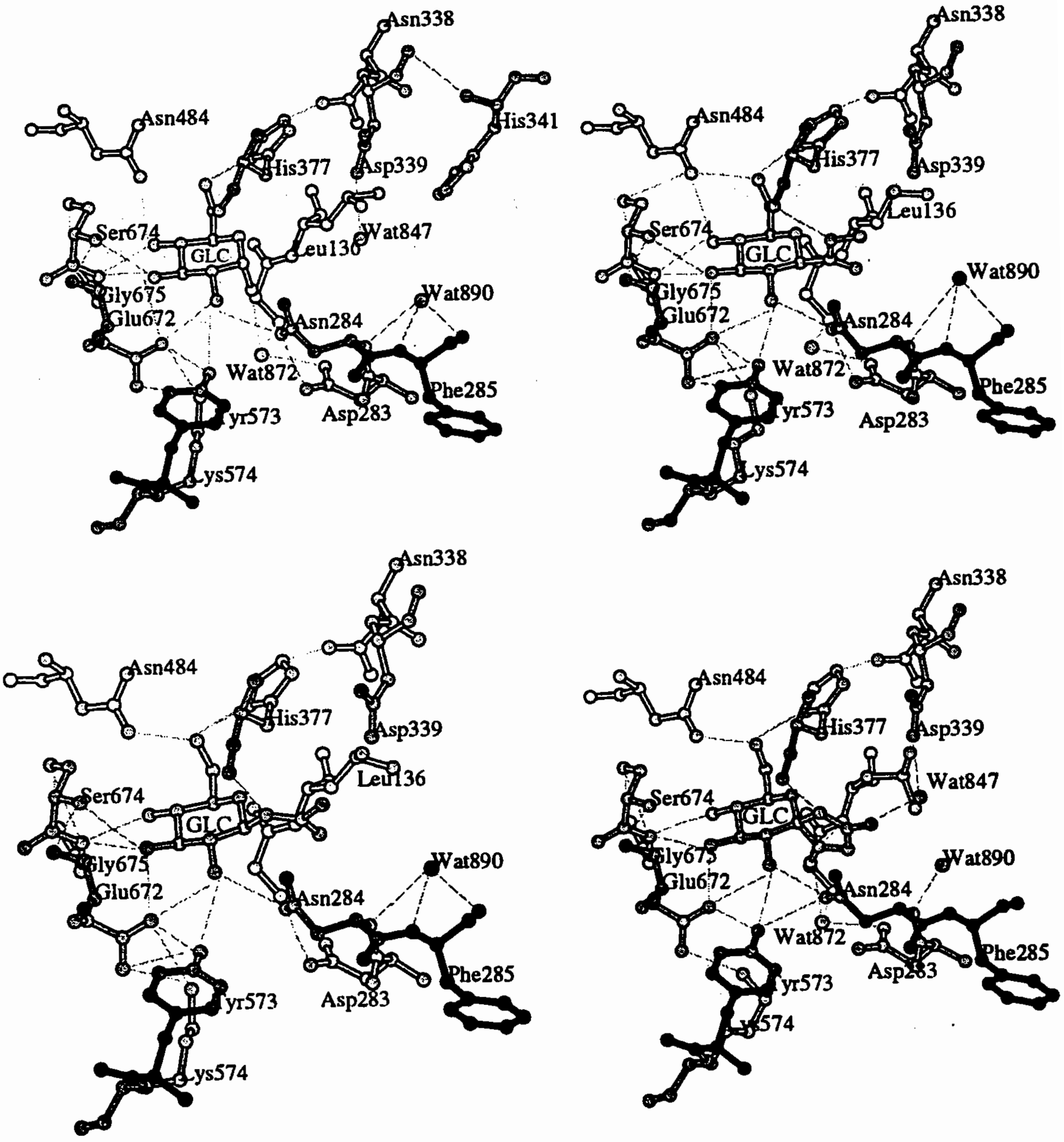


Figure 4. Details of the interactions of glucose analogue inhibitors at the catalytic site of T state glycogen phosphorylase b. a) **1** α-D-glucose b) **5** N-methyl-β-heptonamide c) **6** β-N-acetyl glucopyranosylamine d) **11** the spirohydantoin glucopyranose.

overall slight tightening of the site in which shortening of the hydrogen bonds of the peripheral hydroxyl group on the sugar to the enzyme are noted (Johnson et al., 1994; Watson et al., 1995).

(vii) *Replacement of a non-polar group with a polar group is not always advantageous even in a polar environment.* The methyl group of **6** is 2.8 Å from the position of the water OH4 847 which it displaces. This suggested that a polar group in this position might hydrogen bond to this water and create additional binding energy. Compound **7** has an NH_2 group in place of the methyl but exhibited a poorer K_i of 0.14 mM. The structure of **7** complexed with GPb showed that in order to make the hydrogen bond to the water, the main hydrogen bonding contact to CO His 377 had been lengthened beyond 3.3 Å. The poorer K_i could be accounted for by loss of this hydrogen bond which is not compensated by the hydrogen bond to water. Evidently it is more favourable to displace this water with a non-polar group in the vicinity than to exploit its hydrogen bonding potential (Watson et al., 1995).

(viii) *The change from methyl to ethyl can be neutral.* Compound **8** has an ethyl group in place of the methyl of **6** but exhibited a comparable K_i. Although there may be some favourable energy change from partially shielding the ethyl group on transfer from solvent to the protein, the van der Waals interactions do not provide any substantial gain over those made by the methyl group (Watson et al., 1995).

(ix) *Desolvation effects can be important for polar substituents.* Compound **9**, the N-glycinyl glucopyranosylamine, has an NH_2 group in place of the terminal methyl of **8**, the N-ethyl acetyl glucopyranosylamine. **9** exhibits a higher K_i (0.37 mM vs 0.039 mM). The decrease in affinity may be attributed to a poorer hydrogen bond contact to CO His 377 which is partially compensated by a hydrogen bond OH0 Wat 887 from the carbonyl oxygen and to the energy needed to desolvate **9** on to transfer to the catalytic site. OH4 Wat 847 is displaced but the NH2 group is not able to hydrogen bond directly to Asp339 (Watson et al., 1995).

(x) *Larger groups are not always advantageous.* The complex with N-benzylacetyl glucopyranosylamine **10** shows that the phenyl ring is accommodated in the same region as the CH_2-CH_3 moiety of compound **8** but in order to accommodate the ring the contact to CO His 377 has slightly lengthened(Watson et al., 1995). This compound exhibited a poorer Ki 0.081 mM than the corresponding methyl **6** or ethyl **8** compounds.

(xi) *Rigid groups that are able to exploit several hydrogen bonds are favourable.* Compound **11**, the spirohydantoin, is the best inhibitor to date (Ki 0.003 mM) (Bichard et al., 1995). The NH is able to make the hydrogen bond to CO His377 without distortion (Figure 4d). The CO group hydrogen bonds through water to Asp339 and the α–CO group hydrogen bonds through Wat 872 to Asp 283 (as in the complex with **2**). The other NH group does not make a direct hydrogen bond but is in a favourable electrostatic environment just 4 Å from Asp 283. These extra hydrogen bonds through water appear to compensate the otherwise favourable entropy gain when they are displaced as observed in the complex with **6**. Further we suspect there may be a contribution from the rigid substituent groups, as has been observed in other complexes (Martin et al., 1991) although this has not yet been tested.

Physiology

We have recently explored whether inhibitors, that have been developed on the basis of the rabbit muscle GPb structure, are effective regulators of liver glycogen metabolism. At the time only **6** was available in sufficient quantities. Previous work has shown that **6** (1-GlcNAc) is a competitive inhibitor of both liver and muscle isozymes of GP and is indeed considerably more effective than glucose (Board et al., 1995). In intact hepatocytes 1-GlcNAc has been shown to be an effective regulator of liver GP producing substantial inhibition. At 1 mM concentration 1-GlcNAc promotes activation of liver protein phosphatase by 600 % whereas glucose at 50 mM concentration produces only a 200% enhancement. These results are fully in accord with the regulatory role of glucose, as discussed above and indicate that the glucose analogue inhibitor is considerably more effective than glucose (Board et al., in preparation). The effects on GS are more complex. In gel-filtered liver extracts (where ATP has been removed) 1-GlcNAc leads to activation of GS but in intact hepatocytes there was no direct activation of GS. There is evidence to support the notion that in intact hepatocytes, 1-GlcNAc is metabolised by hexokinase to 1-GlcNAc-6-phosphate and that this compound interferes with GS activation. Despite the lack of activation of GS in intact hepatocytes, subsequent work has shown that glycogen deposition is enhanced and glucose uptake stimulated by the 1-GlcNAc inhibition of glycogen degradation in a dose dependent manner (Board & Johnson, in preparation). These effects on isolated hepatocytes indicate a potential hypoglycaemic action and a positive role for the analogues in the treatment of Type II diabetes. Experiments are underway to assess the effects of 1-GlcNAc on glycogen metabolism *in vivo*.

Acknowledgements. This work has been supported by the Oxford Centre for Molecular Sciences and by the EC contract number B 102 CT94 3025. We also wish to acknowledge contributions by others to this project whose names appear on the original papers.

References

Allemany, S. & Cohen, P. FEBS Lett., 198 (1986) 194.
Bichard, C.J.F., Mitchell, E.P., et al. Tetrahedron Letts., Submitted (1995)
Board, M., Hadwen, M. & Johnson, L.N. Eur. J. Biochem., Submitted. (1995)
Bryan, W.P. Biopolymers, 26 (1987) 387.
DeFronzo, R.A. Diabetes, 37 (1988) 667.
Dent, P., Lavoinne, A., et al. Nature, 348 (1990) 302.
Dunitz, J.D. Science, 264 (1994) 670.
Goodford, P.J. J. Med. Chem., 28 (1985) 849.
Hers, H.G. Ann. Rev. Biochem., 45 (1976) 167.
Johnson, L.N., Watson, K.A., et al. (1994). Glucose analogue inhibitors of glycogen phosphorylase. Complex carbohydrates in drug research: Ed. K. Bock and H. Claussen. Copenhagen, Munksgaard. 214.
Lam, P.Y.S., Jadhav, P.K., et al. Science, 263 (1994) 380.
Martin, J.L., Veluraja, K., et al. Biochemistry, 30 (1991) 10101.
Martin, J.L., Withers, S.G. & Johnson, L.N. Biochemistry, 29 (1990) 10745.
Montgomery, J.A., Niwas, S., et al. J. Med. Chem., 36 (1993) 55.
Perutz, M.F. (1992). Protein Structure: New approaches to disease and therapy. New York, W. H. Freeman.
Reich, S.H., Fuhry, M.A.M., et al. J. Med. Chem., 35 (1992) 847.
Sprang, S.R., Goldsmith, E.J., et al. Biochemistry, 21 (1982) 5364.
Stalmans, W. Curr. Top. Cell Regul., 11 (1976) 51.
Vagelos, P.R. Science, 252 (1991) 1080.
Von Itzstein, M., Wu, W.-Y., et al. Nature, 363 (1993) 418.
Watson, K.A., Mitchell, E.P., et al. Acta Cryst. D, In press (1995)
Watson, K.A., Mitchell, E.P., et al. Biochemistry, 33 (1994) 5745.
Watson, K.A., Mitchell, E.P., et al. J. Chem. Soc. Chem. Commun., (1993) 654.
Wera, S., Bollen, M. & Stalmans, W. J. Biol. Chem., 266 (1991) 339.
Witters, L.A. & Avruch, J. Biochemistry, 17 (1978) 406.

Recent Advances in Automated Ligand Design

Hans-Joachim Böhm
BASF AG, Hauptlaboratorium, D-67056 Ludwigshafen

1. Introduction

A number of computer programs have recently been proposed that attempt to design automatically new ligands for a given protein structure [1-18]. Most programs for the de-novo construction try to assemble novel molecules from pieces. These pieces can be either atoms or larger, chemically reasonable fragments. Atom-based de-novo design programs are LEGEND [2,3] and GROWMOL [4]. Fragment-based programs are GROW [5], NEWLEAD [6], GROUPBUILD [7], HOOK [8] and LUDI [17,18]. Both approaches have advantages and disadvantages. Clearly, the use of single atoms as building blocks will generate the largest possible diversity of chemical structures. All possible structures can be generated by assembling atoms. On the other hand, a large diversity of generated structures can also be obtained with a fragment-based approach by using a large number of different fragments. A potential advantage of the fragment-based approach as compared to the atom-based approach concerns the assessment of the synthetic accessibility of the generated structures. The use of fragments offers the advantage that chemical knowledge can be built into the fragment connection step. For example, amino acids can be used as building blocks to construct peptides [5]. The extension to other simple chemical reactions, e.g. the ether formation, is straightforward. In contrast, this control of synthetic accessibility is not possible for an atom-by-atom build-up program. Therefore, the latter approach requires that the synthetic accessibility is checked at the very end of the design cycle. This is a much more complex task than to check whether the formation of a particular bond is synthetically feasible.

Two major possible strategies exist for the fragment-based approach to de-novo ligand design. One can position several fragments independently (or take them from a known ligand structure) and then search for suitable templates that connect these fragments into one molecule. The advantage of this approach is that the individual fragments are placed without any bond constraints and are likely to be at their optimal positions. Furthermore, this strategy has the desired tendency to generate rigid structures. A possible disadvantage is that it may be difficult to find appropriate templates connecting the fragments in a stereochemically and synthetically reasonable way. The alternative is to start at a certain part of the protein binding site, position one fragment and then append additional fragments in a step-by-step build-up procedure. An advantage of this approach is that chemical knowledge can be easily incorporated into the linking step. Therefore, synthetically accessible structures are more likely to be obtained with this approach. The build-up procedure tends to generate flexible structures. It can run into difficulties if a large gap between two separated parts of the binding pocket has to be bridged without the possibility to form extensive specific interactions with the protein in the gap region.

2. The computer program LUDI

LUDI is a fragment-based de-novo design program which can be used both for 3D-database searching and for the automatic construction of novel ligands either through building (step-by-step build-up) or through linking (placement of individual fragment and subsequent connection). As other programs for ligand design, LUDI requires basically three pieces of information: 1) the 3D-structure of the target protein, 2) 3D-structures of putative ligands or fragments for docking or constructing novel molecules, and 3) information about possible favorable interactions between the protein and the ligand.

An important conceptual aspect of LUDI is its ability to tolerate small uncertaincies in the experimentally determined protein geometry. The accuracy of the experimentally determined protein structures is on the order of 0.2-0.4 Å in the atomic positions. This uncertainty in the protein geometry can give rise to large contributions to the protein-ligand interaction energy if calculated with a molecular mechanics force field. In other words, results from force field calculations can only be interpreted quantitatively if a full geometry optimization including all degrees of freedom is carried out. We have decided to refrain from using a force-field-based scoring algorithm in LUDI. Instead, an error-tolerant empirical scoring function was implemented in LUDI [19]. This scoring function takes into account both enthalpic and entropic contributions to binding.

LUDI constructs novel protein ligands by joining molecular fragments. The program positions molecules or new substituents for a given lead into clefts of protein structures (e.g. an active site of an enzyme) in such a way that hydrogen bonds can be formed with the protein and hydrophobic pockets are filled with lipophilic groups. The positioning of fragments with LUDI is completely based on geometric operations and does not involve any force field calculations.

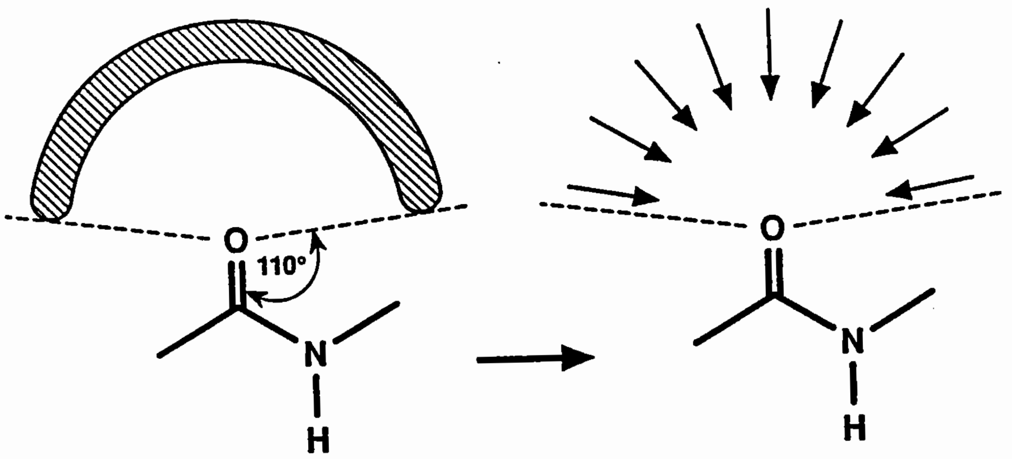


Fig.1 The concept of interaction sites in LUDI. The full range of possible hydrogen bond geometries as evidenced by a statistical analysis of non-bonded contact geometries in crystal packings of organic molecules (shown on the left) is described by a set of interaction sites (shown as vectors on the right)

The program first calculates 'interaction sites', which are discrete positions in space suitable to form hydrogen bonds or to fill a hydrophobic pocket. The interaction sites are derived from a statistical analysis of nonbonded contacts found in the Cambridge Structural Database (CSD) [20,21]. The concept is shown in Fig.1.

The next step is the fit of molecular fragments onto the interaction sites. The fragments are taken from a library. A fragment library containing 1,100 diverse small molecules was generated by us manually using computer graphics. LUDI can also be used to search larger fragment libraries. For example, we use structures from the 'Available Chemicals Directory' (ACD) and the Cambridge Structural Database (CSD) as fragments for LUDI [22]. LUDI can also be run in the 'link-mode'. In this mode LUDI connects some or all of the fitted fragments by bridge fragments to form a single molecule. Alternatively LUDI can append new fragments onto an already positioned fragment or lead compound. In the 'link-mode' LUDI fits fragments onto the interaction sites and simultaneously links them to an already positioned ligand or part of a ligand. The final step is the scoring of the generated protein-ligand complex. The concept of LUDI is summarized in Fig. 2.

When using LUDI, a possible strategy for de-novo design is to carry out first a simple 3D-search (running LUDI in standard-mode) using a 3D-database and select a small number of diverse top scoring hits for biological testing. If experimentally satisfactory binding is observed for some of these structures then they are subsequently submitted to a further LUDI calculation in the link mode searching for substituents. Alternatively, if the 3D-structure of protein complexed with a suitable lead is known, one can use this information to run LUDI in the link mode to search for derivatives.

In a validation study, LUDI was applied to the design of inhibitors of dihydrofolate reductase and HIV-protease [18]. Pisabarro et al. [23] used LUDI in the successful design of novel inhibitors of human synovial fluid phospholipase A_2 with enhanced binding affinity. LUDI can also be used to search large databases of three-dimensional structures for putative ligands of proteins with known three-dimensional structure [22]. As an example, a subset of ≈ 30000 small molecules (with less than 40 atoms and 0-2 rotatable bonds) from the fine chemicals directory (FCD [24]) was used in the search for possible novel ligands for 4 different proteins. The 3D-structures were generated using the program CONCORD [25]. For example, LUDI was applied to the search for ligands for the specificity pocket of the enzyme trypsin. The calculation took only 2 hours on a Silicon Graphics Indigo R4000 workstation and retrieved 153 compounds. The hits obtained with the highest score are p-methyl-benzamidine and benzamidine. Both compounds are indeed found experimentally to bind trypsin with micromolar binding affinity [26,27].

One of the strengths of LUDI is its ability to find small polar ligands for tight polar binding sites as present for example in trypsin. Our current experience indicates that for such proteins the program can retrieve interesting small molecules forming multiple hydrogen bonds with the protein. Due to its commercial availability since 1992 [28], LUDI is now widely used in pharmaceutical industry.

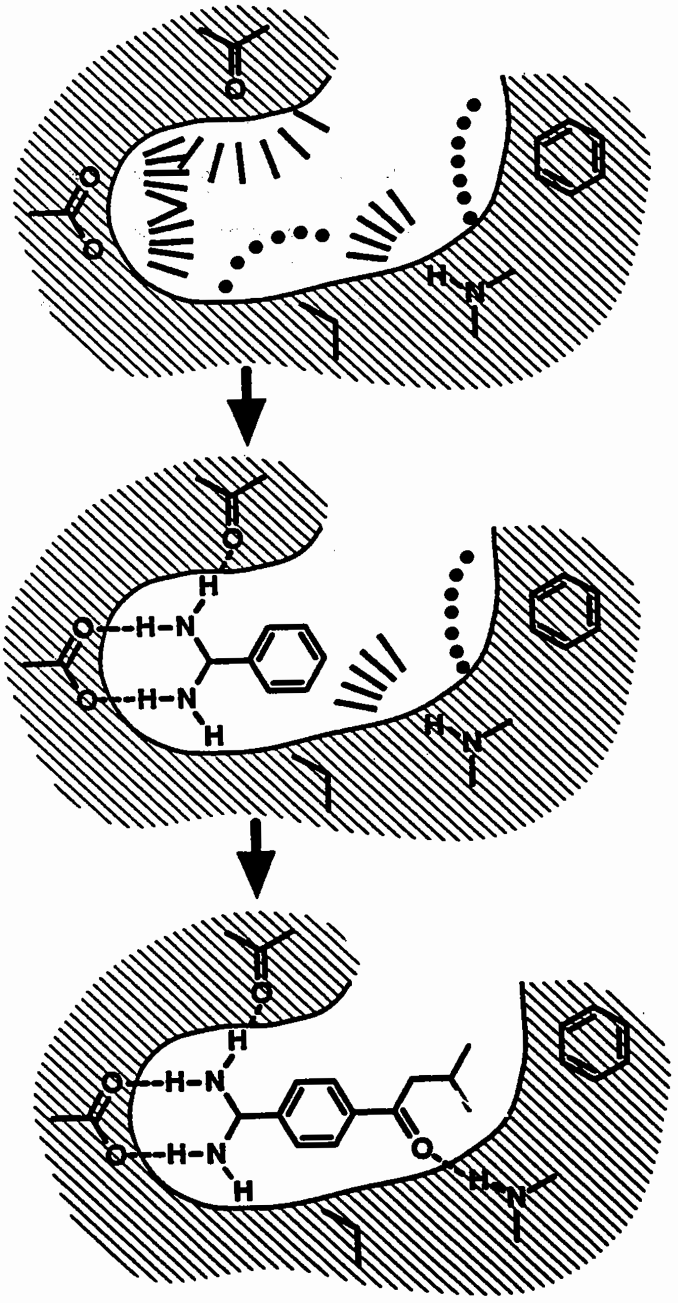


Fig.2. Basis steps of a calculation with LUDI.

3. Limitations of current computational approaches to structure-based ligand design

Despite the remarkable recent progress, the current computational methods for lead discovery face a number of limitations. Most importantly, the programs consider only interactions with the target protein. Transport properties, metabolic stability, or toxicity are not taken into account. It should also be noted that the available current de-novo design programs do not yet address the problem of the synthetic accessibility of the suggested structures. Further, it is clear that current methods for the prediction of the binding affinity need to be improved.

Another important limitation is that in most docking studies and approaches to de-novo ligand design the protein is at present treated as rigid. A number of pairs of 3D-protein structures with and without bound ligand have been determined and give some indication as to which conformational changes of the protein can happen during the binding of a ligand. Some proteins such as trypsin [29] or thrombin [30] have fairly rigid binding sites and do not exhibit large conformational changes upon ligand binding. Sometimes, a part of the protein (e.g. a loop) moves as a consequence of the ligand binding [31-33]. However, it has been shown that this movement is very similar for different ligands. Therefore, even for flexible proteins the 3D-structure of a protein-ligand complex is a good starting point for de-novo design programs.

4. Summary and outlook

The last five years have brought us a dramatic increase in our capabilities to use the 3D-protein structure in the design of novel ligands. Numerous examples for a successful de-novo ligand design have been described [34-42]. Several new computer programs for de-novo ligand design have been described [1-18]. Current experience indicates that LUDI is a useful new aid to the medicinal chemist. Future efforts will focus on incorporating the ability to predict accurately binding energies of putative ligands and coping with the flexibility of both the ligand and the protein.

References

[1] Lewis,R.A. and Leach,A.R., J.Comput.Aided Molec.Des. 8 (1994) 467

[2] Nishibata,Y. and Itai,A., Tetrahedron 47 (1991) 8985

[3] Nishibata,Y. and Itai,A., J.Med.Chem. 36 (1993) 2921

[4] Bohacek,R.S. and McMartin,C., J.Am.Chem.Soc. 116 (1994) 5560

[5] Moon,J.B. and Howe,W.J., Proteins 11 (1991) 314

[6] Tschinke,V. and Cohen,N.C., J.Med.Chem. 36 (1993) 3863

[7] Rotstein,S.H. and Murcko,M.A., J.Med.Chem. 36 (1993) 1700

[8] Eisen,M.B., Wiley,D.C., Karplus,M. and Hubbard,R.E., Proteins 19 (1994) 199

[9] Lewis,R.A., J.Mol.Graphics 10 (1992) 66

[10] Gillet,V.J., Johnson,A.P., Mata,P. and Sike,S., Tetrahedron Comput. Method. 3 (1990) 681

[11] Gillet,V., Johnson,P., Mata,P., Sike,S., Williams,P., J.Comput.Aided.Molec. Design 7 (1993) 127

[12] Rotstein,S.H. and Murcko,M.A., J.Comput.Aided Molec.Des. 7 (1993) 23

[13] Pearlman,D.A. and Murcko,M.A., J.Comput.Chem. 14 (1993) 1184

[14] Lewis,R.A. and Dean,P.M., Proc.Roy..Soc.Lond. B236 (1989) 125

[15] Lewis,R.A. and Dean,P.M., Proc.Roy..Soc.Lond. B236 (1989) 141

[16] Chau,P.L. and Dean,P.M., J.Comput.Aided Molec. Des. 6 (1992) 385

[17] Böhm,H.J., J.Comput.Aided Molec.Des. 6 (1992) 61

[18] Böhm,H.J., J.Comput.Aided Molec.Des. 6 (1992) 593

[19] Böhm,H.J., J.Comput.Aided Molec.Des. 8 (1994) 243

[20] Allen,F.H., Kennard,O. and Taylor,R., Acc.Chem.Res. 16 (1983) 146

[21] Klebe,G., J.Mol.Biol. 237 (1994) 212

[22] Böhm,H.J., J.Comput.Aided Molec.Des. 8 (1994) 623

[23] Pisabarro,M.T., Ortiz,A.R., Palomar,A., Cabre,F., Garcia,L., Wade,R.C., Gago,F., Mauleon,D., Carganico,G., J.Med.Chem. 37 (1994) 337

[24] The fine chemicals directory (FCD) and the available chemicals directory (ACD) are distributed by Molecular Design Ltd., San Leandro, 2132 Farallon Drive, CA 94577

[25] Program CONCORD, distributed by Tripos Ass., 1699 S. Hanley Rd., St.Louis, MO 63144

[26] Mares-Guia,M. and Shaw,E., J.Biol.Chem. 240 (1965) 1579

[27] Recanatini,M., Klein,T., Yang,C.Z., McClarin,J., Langridge,R., Hansch,C., Mol.Pharmac. 29 (1986) 436

[28] LUDI is available from BIOSYM Technologies, 9685 Scranton Road, San Diego, CA 92121-2777

[29] Marquart,M., Walter,J., Deisenhofer,J., Bode,W. and Huber,R., Acta Crystallogr. B39 (1983) 480

[30] Banner,D.W and Hadvary,P., J.Biol.Chem. 266 (1991) 20085

[31] Sali,A., Veerapandian,B., Cooper,J.B., Moss,J.B., Hofmann,T. and Blundell,T.L., Proteins 12 (1992) 158

[32] Rahuel,J., Priestle,J.P. and Grütter,M.G., J.Struct.Biol. 107 (1991) 227

[33] Wierenga,R.K., Noble,M.E.M. and Davenport,R.C., J.Mol.Biol. 224 (1992) 1115

[34] M.A.Navia, M.A.Murcko, Current Opinion in Structural Biology 2 (1992) 202

[35] J.Greer, J.W.Erickson, J.J.Baldwin and M.D.Varney, J.Med.Chem. 37 (1994) 1035

[36] Lam,P.Y.S., Jadhav,P.K., Eyermann,C.J., et al., Science 263 (1994) 380

[37] K.Appelt, R.J.Bacquet, C.A.Bartlett, et al., J.Med.Chem. 34 (1991) 1925

[38] S.H.Reich, M.A.M.Fuhry, D.Nguyen, et al., J.Med.Chem. 35 (1992) 847

[39] Varney,M.D., Marzoni,G.P., Palmer,C.L., et al., J.Med.Chem. 35 (1992) 663

[40] M. von Itzstein, W.Y. Wu, G.B.Kok, et al. Nature 363 (1993) 418

[41] H.Mack, T.Pfeiffer, W.Hornberger, H.J.Böhm, H.W.Höffken, J.Enzyme Inhibition, 1995, in press

[42] B.P.Morgan, D.R.Holland, B.W.Matthews, P.A.Bartlett, J.Am.Chem.Soc. 116 (1994) 3251

Determination, Dynamics and Deviations of the Structure of T4 Lysozyme in 25 Crystal Forms

Xue-jun Zhang and Brian W. Matthews

Institute of Molecular Biology, Howard Hughes Medical Institute and
Department of Physics, University of Oregon, Eugene, OR 97403 USA

Introduction

In the course of characterizing mutants of T4 lysozyme a large number of crystal structures, both in isomorphous and non-isomorphous forms, have been determined (Table 1) (Matthews, 1993; Wozniak et al., 1994; Zhang et al., 1995). These provide an unusually diverse sample to compare the structures of a closely related set of proteins under different crystallization conditions in different crystal packing environments.

Structure Determination Using Modified Rotation and Translation Functions

The crystal structure of wild-type lysozyme, as well as mutants Met 6 → Ile (M6I) and Ile 3 → Pro (I3P) were determined by multiple isomorphous replacement (Matthews & Remington, 1974; Bell et al., 1991; Faber & Matthews, 1990; Dixon et al., 1992). The other mutants were determined by molecular replacement. In some cases, however, the structure determination was complicated by substantial "hinge-bending" within the lysozyme molecule, by the presence of multiple copies of the molecule in the asymmetric unit, or by a combination of both.

In order to deal with such situations methods were developed to enhance the method of molecule replacement by the incorporation of known structural information (Zhang & Matthews, 1994). Generally speaking, this can be done in two ways, either by an "addition" strategy or by a "subtraction" approach. The objective of the former is to increase the signal while the latter is intended to reduce the noise. In addition strategy, the information from the part of the structure that is known is used to supplement the search model for the remaining part still to be solved. The objective is that the structure factors or Patterson function calculated from the enhanced model will better resemble the observed data. For the addition strategy to be effective, the known structural information should provide an independent signal to the correlation function. This requires that the addition term respond to changes of the operator (e.g. a translational operator), rather than being a constant term that is added.

With subtraction strategy, the information from the part of the structure that has been determined is subtracted from the observed data so that the modified observations will more closely represent the part of the structure still to be solved. Use of the subtraction strategy is intended to enhance the desired peaks by reducing noise and by eliminating peaks which correspond to the part of the structure already known. Subtraction strategy can be used only in Patterson space. It cannot be applied in reciprocal space formulations. This is because a calculated structure factor cannot be subtracted from the amplitude of an observed reflection without knowledge of its phase information.

Among a number of possible approaches, the rotation function with subtraction strategy and the correlation translation function with addition strategy were found to be most successful (Zhang & Matthews, 1994). The reason for the former is that the ordinary rotation function includes noise arising from unwanted correlations between different parts of the crystal structure. Subtraction of a part of a structure that is known will delete peaks and noise due to this part, allowing the remainder to become more significant. The reason for the success of the addition strategy in the translation function is that the total model becomes more complete which in turn increases the resemblance of the calculated Patterson function to that observed.

Table 1. Crystal forms of T4 lysozyme

Mutant I.D.	Mutant[a]	Space group	Cell dimensions a(Å) α(°)	 b(Å) β(°)	 c(Å) γ(°)
WT	Wild-type, WT* and many other mutants	$P3_221$	61.2	61.2	96.8 120
M6I	M6I	$P2_12_12_1$	72.2	73.8	150.5
I3P	I3P	$P2_12_12$	86.5	96.6	93.2
S44-[AA]	S44-[AA]/WT*	R32	172.1	172.1	80.0 120
3S-S	I3C/I9C/T21C/C54T/T142C/L164C (Cys3-Cys97;Cys9-Cys164;Cys21-Cys142)	$P2_12_12_1$	63.0	51.0	48.0
I3L	I3L	$P2_12_12_1$	94.2	35.8	56.6
S44E	S44E/WT*	P2	58.1	35.1	46.8 102
R96A	R96A/WT*	C2	115.7	54.8	59.0 103.4
S44F	S44F/WT*	$P2_1$	54.1	55.9	59.9
SS	D127C/R154C/WT* (Cys127-Cys154)	$P4_222$	80.0	80.0	82.3
4008A	T34A/K35A/S36A/P37A	R3	100.5	100.5	40.8 120

6004A	E128A/V131A/N132A/K135A/S136A/ R137A	$P2_12_12$	157.2	177.9	40.5
9001A	E128A/V131A/N132A/K135A/S136A/ R137A/Y139A/N140A/Q141A	$P2_1$	40.4	112.3	135.2 92
T26E	T26E/WT* with covalent peptidoglycan adduct	$P2_12_12$	50.9	67.3	49.6
E45A	E45A/WT*	$P2_12_12$	29.3	129.3	48.9
I3C	I3C (disulfide bridge Cys3-Cys97)	$P2_12_12_1$	97.0	32.0	50.0
QUAD	K16E/K119E/R135E/K147E	$P6_5$	75.1	75.1	54.7 120
2SS	I3C/I9C/C54T/L164C (Cys3-Cys97;Cys9-Cys164)	$P2_12_12_1$	141.2	51.2	47.9
SS-B	D127C/R154C/WT*, Cys127-Cys154, intramolecular S-S bridge	$P42_12$	118.9	118.9	39.0
A146C	C54S/C97S/A146C	$P4_12_12$	53.6	53.6	165.2
T26E-B	Mutant T26E/WT* with non-covalent peptidoglycan	$P2_1$	52.3	57.7	57.6 104.4
SS-C	D127C/R154C/WT*, Cys127-Cys154, intermolecular S-S bridge	$P4_222$	72.6	72.6	82.2
S117V[d]	S117V	$P2_12_12$	147.7	67.1	77.4
J002A	T34A/K35A/S36A/P37A/S38D/N40A/ S44A/E45A/D47A/K48A/WT*	$P6_3$	89.5	89.5	87.1 120
R2	L32T/T34K/K35V/S36D/P37G/S38N/ L39S/WT*	$P2_1$	49.6	127.1	29.1 98.4

---

[a] Mutant M6I, for example, has methionine 6 replaced by isoleucine. Mutant S44-[AA], for example, has two alanines inserted following Ser 44. The designation "WT*" means that the mutant was constructed in the cysteine-free pseudo wild-type lysozyme which includes the substitutions C54T and C97A. 3S-S, for example, is a mutant including three engineered disulfide bridges between residues 3-97, 9-164 and 21-142.

The Influence of Crystal Packing on Structure

One of the questions in macromolecular crystallography is the degree to which crystal packing influences structure, conformation and dynamics.

In order to investigate the lysozyme structures described here for the possible influence of crystal packing it is necessary to take hinge-bending into account. It is also necessary to differentiate structural changes due to crystal packing from those due to the introduction of mutations. Therefore, the lysozyme structures were divided into three

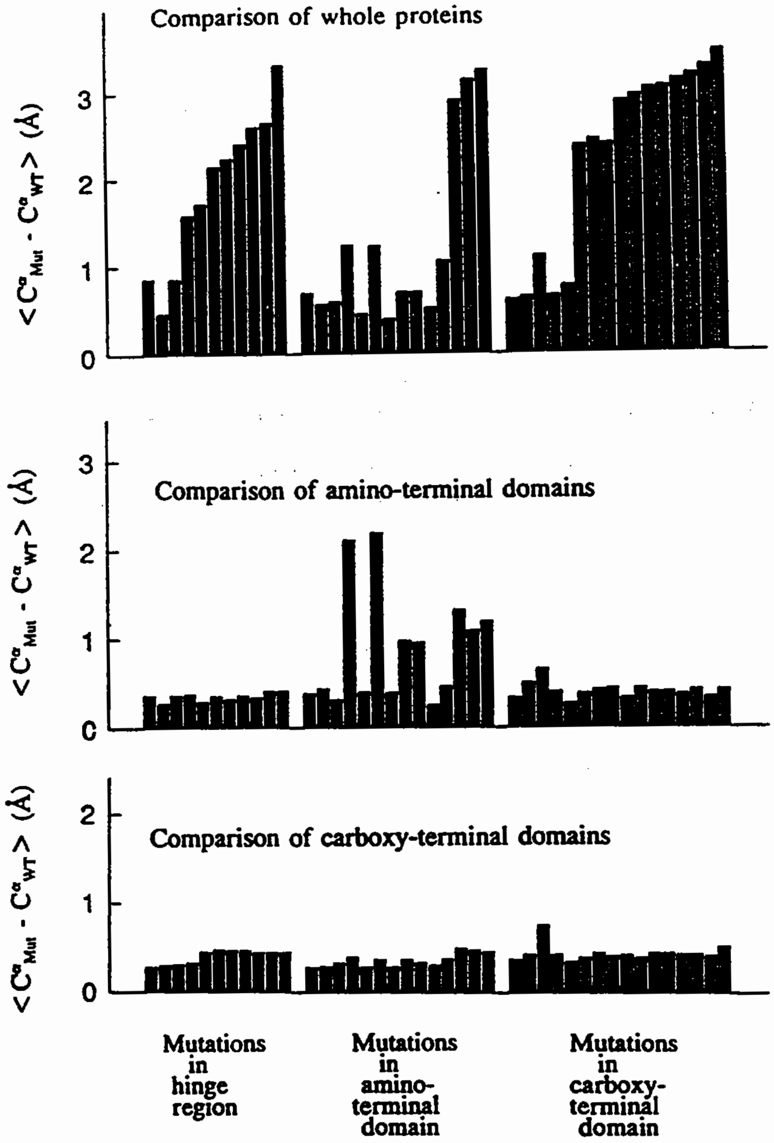


Figure 1(a). (Top panel) Comparison of root-mean-square discrepancy between the α-carbon atoms of each mutant lysozyme (residues 1-162) with wild-type. The lysozyme molecules are separated into three groups. The first group has mutations in the hinge region and, in sequence from the left, with increasing hinge-bending angle, consists of 3SS, $2SS_A$, $M6I_A$, $2SS_B$, I3L, $M6I_B$, $M6I_C$, $I3P_B$, $M6I_D$, $I3P_A$ and I3C. The second group has mutations in the amino-terminal domain and, continuing from left to right, consists of S44E, T26E, E45A, S44-$[AA]_A$, T26E-B_B, S44-$[AA]_B$, T26E-B_A, $R2_A$, $R2_B$, $S44F_A$, $S44F_B$, 4008A, $J002A_B$ and $J002A_A$. The third group has mutations in the carboxy-terminal domain and, continuing from left to right, consists of A146C, SS-C, SS, $R96A_B$, $R96A_A$, $6004A_C$, $9001A_B$, $6004A_A$, $6004A_B$, $9001A_C$, $9001A_A$, $6004A_D$, $9001A_B$, $9001A_D$, $6004A_B$ and SS-B.

Figure 1(b). (Middle panel) Comparisons of the amino-terminal domains (α-carbons 15-59) of mutant lysozymes relative to wild-type (mutants as in Figure 2(a)).

Figure 1(c). (Bottom panel) Comparisons of the carboxy-terminal domains (α-carbons 80-162) of mutant lysozymes relative to wild-type.

groups. (a) Lysozymes with mutations in the hinge-bending region (residues 1-14) and with disulfide bridges across the mouth of the active site. (b) Lysozymes with mutations in the amino-terminal domain (residues 15-59). (c) Lysozymes with mutations in the carboxy-terminal domain (residues 80-162).

Figure 1(a) summarizes the overall correspondence between the α-carbon atoms of each mutant and that of wild-type lysozyme. The discrepancies can be in excess of 3Å, but are dominated by the change in hinge-bending angle of the mutant relative to wild-type.

A more meaningful comparison is to compare the amino- and carboxy-terminal domains separately. These comparisons (Figures 1(b) and 1(c)) allow one to differentiate the effects of the mutations, *per se*, from other effects. Looking first at the amino-terminal domain (Figure 1(b)) some combinations of mutants can lead to root-mean-square backbone changes of 1-2Å. These are, however due either to insertions (S44-[AA]) or to multiple replacements (R2, 4008A and J002A). When the mutations are restricted to the carboxy-terminal domain, the α-carbon shifts in the amino-terminal domain are in the range 0.28-0.66Å and average 0.42Å. (The 0.66Å value corresponds to the disulfide mutant linking Cys 127 and Cys 154.) Conversely, when the carboxy-terminal domains are compared for lysozymes with mutations in the amino-terminal domain (Figure 1(c)) the range is 0.29-0.51Å and the average discrepancy is 0.37Å. The differences do not appear to depend on the resolution to which the structure was determined. Also they do not depend on the hinge-bending angle (Figures 1(b), 1(c)). The discrepancies exceed the estimated coordinate error which, for most of the lysozymes shown in Figure 1, is in the range 0.1-0.2Å. In contrast, the backbone structures of mutant lysozymes crystallized in the *same* (i.e. wild-type) crystal form typically have discrepancies of 0.1-0.2Å, i.e. approximately equal to the estimated error (e.g. Eriksson et al., 1993; Blaber et al., 1994). The discrepancies between the non-isomorphous lysozymes are comparable with those between protein structures determined independently in differing crystalline environments (e.g. Chothia & Lesk, 1986; Wagner et al., 1992; Hohenester & Jansonius, 1994; Kishan et al., 1994). They clearly suggest that crystal packing does influence the structure of the protein to some degree.

The Influence of Crystal Packing on Dynamics

Refined models of crystal structures give not only the coordinates of each atom but also the thermal factor ("B-factor"). The latter provides a measure of the mean square amplitude of displacement or "mobility" of each atom. Similar molecules in different packing environments can retain similar thermal factors (Artymiuk et al., 1979). This suggests that mobility as estimated by crystallographic thermal factors is related to mobility in solution, and that crystal contacts only influence thermal factors to a limited degree. The degree to which the thermal factors for the individual residues vary from structure to structure is illustrated in Figure 2. There is substantial fluctuation from structure to structure although some regions remain consistently more mobile than others. Not surprisingly, the side-chain thermal factors show greater variability than the main-chain. Consistent patterns are, however, apparent, for example the alternation between high mobility and low mobility within the long interdomain helix (residues 60-79), corresponding to residues that are alternatively on the solvent-exposed and the buried sides of the helix.

The consistency of the thermal factors from structure to structure was evaluated by determining the correlation between the thermal factors of all pairs of molecules in the set. The side-chain thermal factors in the various structures (Figure 3) agree substantially better than do those for the main-chain atoms (Figure 3). Presumably this is in part because the thermal factors of the side-chains have greater variability from residue to residue than do the main-chain atoms. Also, although crystal contacts may restrict the motion of some side-chains, most side-chains clearly retain similar thermal factors throughout all the structures. Unexpectedly, there are a number of pairs of molecules for which the correlation coefficient between the main-chain thermal factors is essentially zero, or even negative in some cases. In the most extreme case (I3P$_B$ and the double alanine insertion mutant S44-[AA]$_B$), the correlation coefficient between the main-chain thermal factors is -0.32.

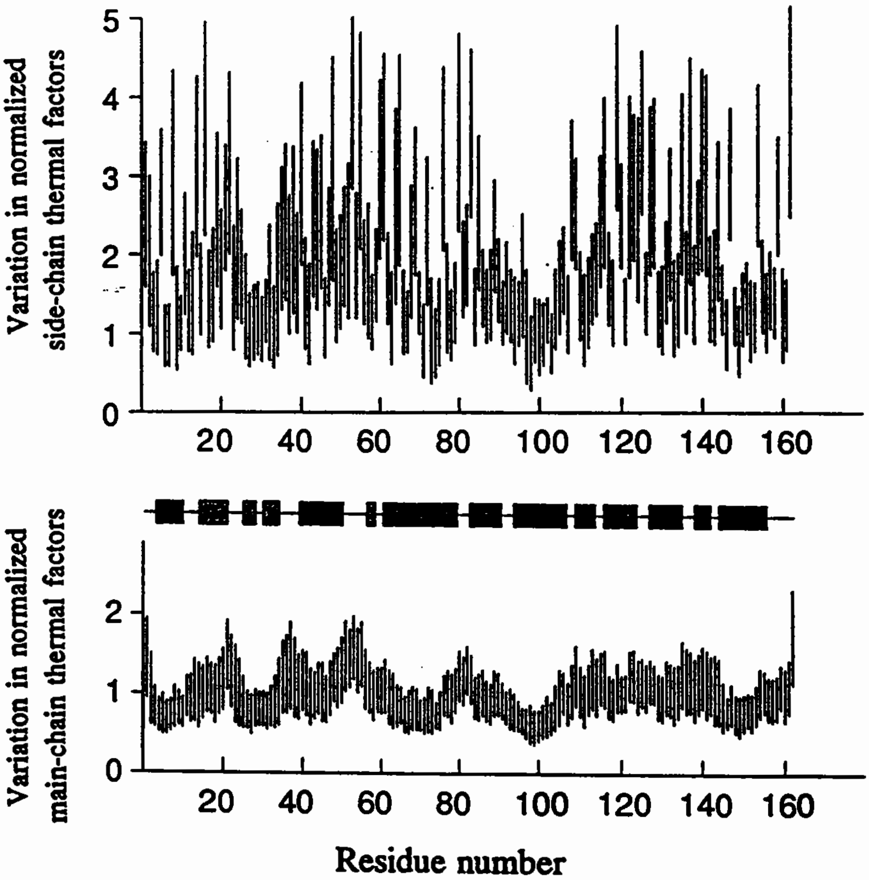


Figure 2. Variability of thermal factors for lysozyme structures in different crystalline environments. The overall thermal factor of each structure was first normalized so that the average value was equal to wild-type. This corresponds to 1.0 in the figure. The average thermal factor $\bar{B}$ and the standard deviation, $\sigma(B)$, were determined for each residue in all the structures. The vertical bars in the figure correspond to $\bar{B} \pm \sigma(B)$. The darker rectangles show the locations of the α-helices and the lighter rectangles indicate β-sheet. Side-chain atoms shown above and main-chain atoms below.

	A146C	3SS	4008A	6004A_A	6004A_B	6004A_C	6004A_D	6004A_E	9001A_A	9001A_B	9001A_C	9001A_D	9001A_E	R2_A	R2_B	2SS_A	2SS_B	SS_C	E45A	I3C	I3L	I3P_A	I3P_B	J002A_A	J002A_B	M6I_A	M6I_B	M6I_C	M6I_D	QUAD	R96A_A	R96A_B	S44E	S44F_A	S44F_B	S44_AA_A	S44_AA_B	SS	SS_B	T26E	T26E_B_A	T26E_B_B	WT
A146C	100	62	32	33	57	32	-47	30	3	14	-1	8	-15	58	58	28	10	74	50	46	58	40	-14	55	43	24	17	54	27	63	38	86	54	37	11	42	75	63	54	65	30	62	62
3SS	72	100	42	34	62	31	-43	25	4	16	-4	8	-14	46	45	19	22	57	55	51	37	32	-16	45	38	30	26	52	31	68	28	44	47	37	8	30	68	44	40	74	33	64	52
4008A	44	52	100	17	38	37	25	21	31	25	22	30	3	40	34	30	23	19	31	40	5	23	3	30	37	35	54	43	38	50	39	48	38	34	33	30	44	19	44	54	51	52	36
6004A_A	46	54	28	100	38	40	18	58	28	38	-2	30	29	-6	3	27	22	18	11	23	-1	11	1	20	41	38	20	38	29	47	30	32	43	38	11	34	36	14	45	30	43	43	27
6004A_B	68	63	44	64	100	56	32	53	33	27	34	29	18	46	45	43	15	41	38	54	25	38	11	48	31	38	43	57	47	58	58	58	53	48	33	51	60	34	61	50	57	68	64
6004A_C	52	67	46	54	70	100	43	56	43	19	52	40	21	35	30	31	27	18	28	40	18	13	13	32	37	32	54	46	46	55	48	54	38	31	46	31	30	13	52	38	65	45	36
6004A_D	27	32	35	38	57	62	100	48	58	29	51	48	38	11	8	51	12	-29	-1	17	-26	8	61	15	-8	50	54	15	61	-1	65	35	28	39	58	46	-17	-13	32	-61	41	17	7
6004A_E	55	60	38	61	75	70	62	100	42	32	29	34	37	18	15	37	33	15	13	35	12	14	30	29	38	53	38	35	60	48	48	38	46	44	34	33	17	11	58	30	51	47	39
9001A_A	44	48	42	48	53	55	54	68	100	16	38	34	18	7	1	19	23	-19	18	20	-14	7	17	10	17	41	44	33	37	28	33	22	20	28	28	23	12	-16	35	10	43	25	8
9001A_B	38	40	28	55	46	38	32	50	38	100	30	19	18	-4	0	25	14	-8	12	11	-18	4	14	-1	1	32	27	25	32	24	34	34	24	35	18	29	12	-8	23	18	19	25	11
9001A_C	38	39	33	30	58	70	53	52	47	38	100	34	20	19	8	18	7	-15	10	28	-7	-8	29	8	3	19	49	29	38	14	33	27	18	13	37	17	-4	-11	22	5	41	11	5
9001A_D	48	46	38	53	58	58	52	60	58	38	48	100	34	4	2	33	19	-6	8	31	-8	14	23	27	19	53	43	34	60	38	46	38	28	40	29	34	5	8	39	19	41	27	27
9001A_E	35	37	24	51	54	58	54	57	48	38	58	52	100	-20	-14	21	1	-19	-13	11	-19	-1	19	8	19	27	25	8	38	7	20	11	11	18	17	29	-13	-7	28	-8	34	13	3
R2_A	68	57	40	32	67	58	41	47	41	28	40	40	43	100	69	30	13	70	57	51	62	40	-6	58	34	1	34	29	19	44	46	85	32	32	34	40	50	65	38	44	40	38	67
R2_B	72	60	41	31	67	58	37	48	40	28	43	41	42	83	100	33	14	60	48	48	72	51	-11	63	38	-4	23	25	18	48	44	62	32	31	38	38	53	78	35	44	32	38	64
2SS_A	58	68	47	52	62	58	57	63	58	46	42	52	48	61	59	100	2	23	19	28	7	32	34	41	19	41	38	28	53	28	64	44	54	50	39	53	15	34	37	11	41	42	51
2SS_B	48	60	40	47	65	58	48	63	50	35	37	44	58	58	58	74	100	3	9	46	19	8	10	7	38	28	37	21	18	38	14	19	14	11	21	-4	11	1	21	23	33	27	20
SS_C	60	58	32	40	62	46	18	52	35	29	28	43	34	73	78	67	47	100	42	38	77	51	-34	57	47	-10	-1	30	2	58	34	38	33	28	3	28	63	83	35	53	23	40	68
E45A	63	57	38	44	58	53	33	48	38	38	38	40	33	72	68	68	52	63	100	41	35	32	-4	54	28	18	23	38	34	39	33	33	28	32	5	28	39	35	32	48	31	38	50
I3C	68	68	46	50	68	64	48	63	48	41	58	58	57	65	62	67	64	58	61	100	38	43	8	58	44	31	58	57	48	55	38	43	40	35	34	35	46	42	48	64	57	54	48
I3L	72	62	33	37	63	53	32	57	48	31	38	42	42	74	79	62	61	78	67	68	100	43	-13	58	46	-14	0	18	-8	40	13	37	18	13	19	8	37	68	35	38	17	27	50
I3P_A	52	50	38	34	58	53	48	54	40	34	38	48	38	58	65	52	46	58	52	53	59	100	5	51	35	35	22	40	25	28	38	41	41	38	25	47	48	65	43	32	29	38	44
I3P_B	27	34	30	25	41	40	53	48	38	38	53	36	33	27	28	54	47	20	38	41	38	33	100	5	-4	48	25	13	38	-16	48	21	31	32	42	29	-42	-6	20	-29	16	13	19
J002A_A	61	68	38	50	61	58	51	62	44	34	40	53	52	68	68	71	65	58	64	72	71	58	47	100	63	20	38	41	37	49	46	55	42	48	38	48	40	58	58	43	46	48	55
J002A_B	57	58	42	51	48	53	37	53	46	38	33	46	43	53	55	61	61	53	58	67	68	48	42	84	100	14	38	43	17	54	21	48	35	34	21	21	47	41	48	50	51	48	46
M6I_A	55	62	49	52	58	62	57	62	55	46	45	51	52	58	48	71	65	42	54	68	51	46	52	58	58	100	58	44	68	27	58	38	48	65	55	44	16	2	48	17	38	44	20
M6I_B	52	58	64	34	55	63	57	54	51	32	54	41	46	60	58	62	58	38	50	68	54	60	47	58	57	64	100	52	58	42	51	65	37	38	53	35	38	7	48	38	63	43	28
M6I_C	65	68	58	48	68	57	44	58	64	34	47	47	48	58	58	62	65	54	58	70	65	47	48	63	65	71	74	100	42	58	41	67	58	41	38	42	54	27	52	58	47	50	46
M6I_D	54	61	45	46	65	57	67	68	58	34	48	52	54	46	58	64	58	41	47	68	53	46	54	58	51	75	73	67	100	46	67	48	52	52	48	55	14	14	43	22	54	44	42
QUAD	68	77	52	46	61	68	35	68	58	32	38	58	33	60	60	53	62	62	60	67	65	53	38	60	57	58	62	65	62	100	33	51	48	42	18	23	57	38	48	71	62	62	61
R96A_A	63	62	48	46	60	61	63	68	58	42	51	50	61	67	67	78	64	58	55	68	62	50	51	68	58	75	60	67	73	63	100	61	55	65	53	64	21	40	53	20	51	43	58
R96A_B	67	63	53	46	65	62	58	60	52	38	52	48	52	65	67	74	62	58	60	71	67	58	46	78	68	65	75	78	64	60	60	100	58	38	62	58	47	43	68	40	54	61	50
S44E	70	67	44	65	62	57	61	68	54	52	43	54	51	54	58	74	57	60	52	61	58	50	44	68	65	67	58	62	65	59	72	72	100	65	38	58	47	37	51	35	48	63	58
S44F_A	64	68	43	58	65	58	53	68	51	45	43	63	51	58	58	61	65	58	68	72	65	53	51	72	65	73	57	67	68	63	70	74	74	100	25	48	28	37	41	38	47	48	53
S44F_B	48	53	44	43	55	60	64	63	53	38	40	51	48	58	58	70	58	48	55	58	62	57	51	68	63	60	68	60	65	55	72	78	63	68	100	44	11	18	43	-1	51	28	28
S44_AA_A	68	60	46	46	65	54	58	65	53	42	43	48	48	64	62	67	51	57	51	58	58	50	41	58	48	65	62	63	63	51	70	70	70	67	61	100	42	42	58	19	38	44	38
S44_AA_B	78	67	40	47	58	46	18	48	38	31	30	34	29	64	63	53	41	70	60	60	59	48	20	55	58	40	49	61	38	55	54	58	65	58	48	63	100	52	43	73	40	55	47
SS	72	58	38	46	60	51	33	58	40	38	33	54	46	67	70	68	53	85	65	61	72	58	29	71	58	47	45	50	52	59	65	64	62	68	60	59	60	100	39	46	22	33	65
SS_B	70	62	46	57	73	73	54	70	58	43	52	58	58	64	64	78	67	67	62	70	68	58	38	73	65	85	64	65	62	67	74	73	68	71	68	68	58	73	100	33	38	51	37
T26E	68	68	55	47	85	45	33	43	40	40	38	49	38	57	61	58	53	60	58	71	57	46	18	58	60	52	55	68	47	59	85	62	62	61	47	52	68	50	62	100	58	58	53
T26E_B_A	57	67	43	57	63	60	52	65	53	42	48	65	51	58	58	73	71	51	50	68	65	50	45	69	65	63	61	64	64	64	68	68	67	73	65	58	52	61	67	63	100	61	51
T26E_B_B	64	72	48	50	64	58	43	60	65	44	44	65	40	55	59	73	67	57	60	68	68	50	38	68	64	62	58	64	58	64	67	68	68	72	58	60	58	65	72	70	68	100	58
WT	68	68	41	58	64	54	40	65	48	46	43	58	52	62	65	71	58	70	64	70	67	62	38	73	65	63	51	67	61	59	70	71	71	74	61	58	58	77	68	67	68	71	100

Figure 3. Matrix showing the correlation coefficients between the thermal factors of each mutant lysozyme with all other molecules in the set. Mutants are identified as in Table 1. The upper diagonal half of the matrix shows the correlations for the main-chain thermal factors. The lower diagonal shows the correlations for the side-chain thermal factors.

Overall, the generally high correlations between the side-chain thermal factors (Figure 3 is reassuring. This tends to confirm that these thermal factors are a reliable indication of side-chain mobility in solution. On the other hand the low or negative correlations in some cases for main-chain thermal factors (Figure 3) shows that these must be used with caution. Indeed, in some cases main-chain motion, as judged by thermal factors, appears to be constrained to such a degree by intermolecular contacts that it is no longer indicative of motion in solution. Tightly packed crystal structures tend to diffract to high resolution and so give more accurate thermal factors. At the same time these are the very structures where the main-chain thermal factors may be constrained by crystal contacts and be so less representative of dynamic behavior in solution.

Acknowledgements

We are especially grateful to Sheila Snow and Joan Wozniak for purifying and crystallizing many of the mutant lysozymes described here. This work was supported in part by NIH grant GM21967 to B.W.M.

References

Artymiuk, P.J., Blake, C.C.F., Grace, D.E.P., Oatley, S.J., Phillips, D.C. and Sternberg, M.J.E. Nature, *280* (1979) 563
Bell, J.A., Wilson, K.P., Zhang, X-J., Faber, H.R., Nicholson, H. and Matthews, B.W. Prot: Struct Funct Genet, *10* (1991) 10
Blaber, M., Zhang, X-J., Lindstrom, J.D., Pepiot, S.D., Baase, W.A. and Matthews, B.W. J Mol Biol, *235* (1994) 600
Chothia, C. and Lesk, A.M. EMBO J, *5* (1986) 823
Dixon, M.M., Nicholson, H., Shewchuk, L., Baase, W.A. and Matthews, B.W. J Mol Biol, *227* (1992) 917
Eriksson, A.E., Baase, W.A. and Matthews, B.W. J Mol Biol, *229* (1993) 747
Faber, H.R. and Matthews, B.W. Nature, *348* (1990) 263
Hohenester, E. and Jansonius, J.N. J Mol Biol, *236* (1994) 963
Kishan, K.V.R., Zeelen, J.P., Noble, M.E.M., Borchert, T.V. and Wierenga, R.K. Prot Sci, *3* (1994) 779
Matthews, B.W. Ann Rev Biochem, *62* (1993) 139
Matthews, B.W. and Remington, S.J. Proc Natl Acad Sci USA, *71* (1974) 4178
Wagner, G., Hyberts, S.G. and Havel, T.F. Ann Rev Biophys Biomol Struct, *21* (1992) 167
Wozniak, J.A., Zhang, X-J., Weaver, L.H. and Matthews, B.W. Molecular Biology of Bacteriophage T4 (J.D. Karam, ed.) (1994) 332
Zhang, X-J. and Matthews, B.W. Acta Cryst, *D50* (1994) 675
Zhang, X-J., Wozniak, J.A. and Matthews, B.W. Unpublished results (1995)