A Semi-Automated Map Fitting Procedure

T.J.Oldfield
Molecular Simulations Inc, 9685 Scranton Road, San Diego, CA 92121, USA.
and
Dept. of Chemistry, University of York, Heslington, York, YO1 4LD, UK
Email: tom@yorvic.york.ac.uk
URL: http://www.yorvic.york.ac.uk/~tom

Abstract

The electron density applications available within QUANTA96 represent novel and effective tools for speeding up the processes of C_α tracing experimental maps and manual rebuilding during refinement. The various modules (X-AUTOFIT, X-BUILD, X-SOLVATE and X-LIGAND), have been developed over the past year in close collaboration with the large number of crystallographers working on projects in the Protein Group at York. Crucially, these new tools are easy to learn, and natural to use, and provide a 10 fold reduction in the time spent at the graphics terminal.

1 Introduction

The determination of the detailed three dimensional structure of a protein molecule by X-ray crystallography involves a number of key steps. These include the crystallisation of the protein, data collection, data processing, phase determination, model building, refinement and analysis of the final coordinates. This paper presents some novel procedures and algorithms that can greatly reduce the time taken for both the initial tracing during the model building stage and for rebuilding during refinement. They are incorporated in the programs X-AUTOFIT, and X-BUILD which are embedded in QUANTA.

The interpretability of the map is sensitive to the quality of the initial phases and there is still no definitive measure of this. The best indicator is based on map skeletonisation. In 1974 J. Greer described an algorithm that carries out this skeletonisation by charting a line through the electron density along the direction of minimum descent [1]. These lines are commonly called bones, and are often flagged as a series of coordinate points sometimes referred to as the bones coordinates. They should indicate the probable connectivity of the protein. The program O [2] is now the standard program that uses a bones skeleton, generated by an external program, to guide the initial fitting of C_α coordinates into electron density. X-AUTOFIT aims to provide an application where skeletonisation of the density is integral to the working of the program, and as much help as possible is provided by semi automated map fitting tools which use this skeleton.

X-AUTOFIT and X-BUILD are not designed to replace experienced crystallographers, but rather provide them with tools that allow map fitting to proceed faster, and to assess the results at each stage against prior knowledge. The program has been developed in response to the requirements of the crystallographers working in York, and is under ongoing development. While automation is possible with a perfectly phased map, maps based on experimental phases are not perfect, and indeed the quality of the electron density varies over the map itself. All the algorithms have been designed to work in real time; users become impatient if they have to sit idle while their precious graphics time seeps away.

2 New techniques

The new algorithms and techniques described here cover the following aspects of crystallographic determination of

protein structures; map masking, chain tracing electron density, sequence assignment, general model building, real space refinement, solvent fitting, ligand fitting and validation of structure. All calculations and functionality are integral to the program and can be used in any combination. The theory will be described in this section, and details of the user interface and practical applications discussed in section 4.

2.1 Map Masks

A map mask can be either calculated from a bones skeleton (see following section on the use of bones), from a set of atomic coordinates, or read from a file. This mask can be saved to a file for use by other external programs such as DM [3]. The main advantage of X-AUTOFIT map masking is the ease with which the actual mask can be manipulated. Voids within the mask can be automatically deleted using a fast void searching algorithm [4] to leave a clean simple surface mask. The actual extent of the mask can be edited interactively using a moveable spherical pointer with a changeable radius. It is possible to add or remove mask volume with the cursor, and since the recalculation of the surfaces is so fast, any changes are instantly observed. As X-AUTOFIT was developed on a modest power computer, it soon became obvious that the actual size of the mask was a problem for general manipulation. This problem has been solved using two approaches. First the mask is drawn as a dot surface rather than as a net to reduce the size of the graphical object. Secondly, the number of points in the net can be artifically reduced (and increased) with a single tool thereby making its size much more manageable for use on smaller graphical machines. Since the mask can be readily manipulated, as well as turned on and off, it is possible to make use of it in other building processes. The most obvious is during the initial the C_α-tracing as at this stage, symmetry related atoms may not yet be built at a crystal contact. Here the mask can be used as a graphical bounding surface.

2.2 Electron density skeletonisation

The skeletonisation of density is a three dimensional data reduction algorithm that removes points from the electron density using the four rules described in the orginal paper of J Greer [1]. Mathematically this is simple to implement, but difficult to make fast enough to provide an interactive display. With careful consideration of memory and the use of short cuts in the calculation it has been possible to carry out the calculation on 100,000 map points (nominally a 12Å radius sphere of electron density) per second (Indigo R4000). This allows the user to change parameters in the algorithm via a user dial and watching the effect on the resultant electron density skeleton. This is particularly useful when the chain traverses the core of the protein where there is good density, then forms a loop at the surface, where lower density levels must be used to observe chain connectivity.

The actual speed of the algorithm for reducing the map to a skeleton in comparison to that of O bones program is difficult to define; the speed is sensitive to the sparsity of the map and the cache size of the computer. Although there is about a five fold improvement in the calculation times; the user perceives a larger apparent increase in speed because the skeleton is only generated around the place of interest in the map. The algorithm tends to proceed in an indeterminant way, and it is difficult to know the best initial density level to use. The user can observe the effect of changing this and hence make a better informed choice in the building process. Some changes have also been made to the algorithm to improve the connectivity by a different treatment of neighbours in the map. This results in a more connected skeleton. A fast skeleton smoothing function is included. The user has the option to draw a bicubic spline curve through the points of the skeleton to give completely smooth bones. The curves are arranged so as to retain the position of the branch points and end points in the skeleton and therefore keep the important features used to guide building in the correct positions.

A tree search function processes the skeleton to provide automatic determination of side chain and main chain bones. This function is also used to trim the isolated fragments to give a cleaner representation of the density. The depth of search used to determine side chain and main chain segments can be changed to tune the algorithm to a particular map. It is also possible to reassign bones from main chain to side chain or to delete sections with a single point and click action. The main chain and side chain bones are coloured differently.

A function is also provided that endeavours to assess the quality of the map by a quantitative measure of the number of breaks and false links in the skeleton. In a perfect map there should only be a single chain for each segment, possibly linked by disulphide bridges, and all the density will be attached to this single tree. Therefore the quality of the map is in some way proportional to the number of separate trees and false links. This function can be used to screen several maps produced using different phasing methods (histogram matching, a variety of solvent flattening techniques, and so on) and to suggest the most interpretable map.

2.3 Semi automated C_α tracing

The first step in building a new structure is to trace the polypeptide chain through the electron density. A novel algorithm has been designed to aid in this building process that makes use of the bones and the density gradient to provide suggestions for the next C_α position in the chain trace.

The C_α atom is placed using a series of rules. The algorithm considers:

1) The set of positions which are linked to the previous C_α atom by continuous skeleton and 3.8Å away from the previous C_α atom.

2) The C_α geometry with respect to the previous C_α atoms fitted, weighted as a function of the C_α conformation map. It has been shown that a probability surface can be generated if the pseudo torsion from each set of four C_α atoms in a protein is plotted against the pseudo opening angle derived from three C_α atoms (subsets of same 4 C_α atoms) for a well refined set of proteins from the Protein Data Bank [5]. The allowed regions on this plot are restricted. Different areas on the plot provide information on the local conformation of the protein backbone, such as alpha helix, beta sheet, and turn structures. It is therefore possible to use this empirically derived probability surface to direct the fitting of C_α atoms to electron density.

3) The analysis of branches at/near the set of points 3.8Å from the last C_α atom.

4) Whether the bones segment represents a side chain or main chain section of electron density.

5) Real space refinement of the pseudo angle, torsion and bond length around the local area found by bones analysis.

The program places the next C_α atom at a position which has the greatest weight based on the five rules described. The probability of the program determining the correct position depends on the quality of the map, but ranges from 90% for a 2fo-fc map (test case), to 40% for a MIR map of moderate quality [6]. Since the algorithm is weighted towards observed protein conformations, it allows the user to build secondary structure elements very quickly . It can often recognise that a particular connectivity may be the result of side chain density merging into another section of the map; this is particularly important when building beta sheets. This rule based building algorithm allows large parts of the structure to be built quickly and with confidence.

There are also tools for the general manipulation of the C_α trace such as joining and cutting the trace during the building process, and for positioning single C_α atoms. Multiple C_α atoms can be generated in predefined helix and strand conformations, and any part of the trace can then be rigid body refined into electron density.

2.4 The C_α-plot and the Ramachandran plot

The program provides the empirical probability surface contoured to highlight various conformations common in proteins. A pointer on this map indicates the current value of the angle and torsion, and its position can indicate the occurrence of helix or extended conformations. Hence, it is immediately obvious when the C_α atoms are being fitted to helix, or beta-strand conformations, or when impossible conformations are being generated, such as when the C_α trace is extended along side chain electron density. This plot also has the advantage of indicating when the user is generating extensive left handed helices into their maps - a feature that suggests that the wrong hand for the heavy atom solution has been used to phase the map. An added advantage of the C_α conformation map is that the user can pick a point from this map, and so place the current C_α atom into this conformation, for example, an alpha helix.

A Ramachandran plot is displayed during the model building procedure where all changes in coordinates are reflected by changes in the distribution of the Ramachandran points on a contoured plot. This plot can also be picked with a cursor so that a residue with poor angles can be identified, and futher more, the program will automatically centre on this residue that has been selected.

2.5 Sequence assignment by fuzzy logic

At some stage in the building process the crystallographer has to fit the known sequence to this polypeptide fragment. The only information available is the extent of the map density at each C_α coordinate fitted to the map. Inspection of the electron density will occasionally indicate the position of tryptophans or prolines, but in

most cases the density may only indicate size or the environment . The sequence alignment algorithm presented here allows the crystallographer to define a residue type by a fuzzy descriptor. The present fuzzy descriptors are: Unknown, Big, Medium, Small, Aliphatic, Aromatic, Polar, Non-polar, Charged, Acid, and Basic. Each of the 20 common residues found in proteins have a propensity likelihood associated with each of the fuzzy definitions. These can be assigned a linear scale of likelihood ranging from 0 to 10. For example glycine rates 10/10 for small, 1/10 big, and 0/10 aromatic. A value of zero represents complete exclusion in the alignment. A known residue is given a propensity value of 10, and all other residue types return zero from the alignment at this position. The user can change the weighting of each of the descriptors by supplying their own table of propensity values. For example, it is possible to define lysine as both Big and Small because the side chains are sometimes not visible in density. In this case they the propensity value for lysine might be set to 8/10 Big and 8/10 Small.

Once the user assigns fuzzy descriptors to the residues in a fragment the program shows on a sequence table all the possible forward and backwards sequence aligments that match the sequence to the C_α definitions. The alignments are shown as blue arrows for forward fits and red arrows for backwards fits, while the thickness of the arrow indicates the goodness of fit. The program only shows sequence alignments where all the residue matches are non-zero, and the sum of the fuzzy propensities is greater than five. An unknown residue type takes no part in the alignment.

The number of possible hits is indicated by the number of alignment arrows, while the respective thickness of the hit arrows will narrow down the possible section of the sequence that corresponds to the fitted C_α atoms.

The fuzzy sequence alignment algorithm is instantaneous, so the crystallographer can experiment with different descriptions of residue types and watch the effect on the alignment.

2.6 Amino acid main chain and side chain density fitting

The program has a torsion angle real space refinement procedure that will build entire residues from the C_α atoms. The real space refinement procedure first fits polyglycine to the C_α coordinates. It then checks the geometry of the fitted residues to find which residues have been built in a conformation which results in poor main chain angles. The algorithm then uses the correct sections of residues to refine the bad residue conformations as consecutive residues have correlated angle information. In this way the program can produce a reasonable polyglycine chain from an apparently poor initial trace.

Once the backbone atoms have been fitted, the C_β atoms are positioned to give tetrahedral L amino acid C_αs. The program then searches for the side chain by wobbling the C_β atom to find the best direction. This is necessary as errors in the placement of the C_α can result in rigid refinement procedures missing the side chain density altogether. The remaining side chain atoms are fitted one by one using chi angle refinement, then all atoms are fitted together to the density. Again the program allows a wobble for each chi angle so that the side chain can follow density even when the C_α coordinate has an error in the region of 1Å .

This procedure has a very large radius of convergence, but will produce a model with coordinate errors of up to 0.5Å. These are within the radius of convergence of Xray reciprocal space refinement techniques. The aim of the real space refinement algorithm described here is to find and fit side chain atoms tolerating some geometric errors which can easily be corrected later.

Proline residues are treated differently. Once the program has fitted the backbone atoms for a proline residue, the C_β, C_γ and C_δ atoms are placed by refinement of the torsion about the C_α-C_β bond. The program also checks that the internal geometry of the proline is within reasonable limit during the refinement. Both cis and trans conformations are tested, and the program will select the configuration which gives the best fit to the density for this residue.

The coodinates generated by this algorithm are coloured according to how well they fit the density so that poor sections of the build are flagged.

2.7 Checking the chain direction

The real space refinement algorithm can also be used to check whether the C_α chain is built in the correct direction, and also whether C_α positions have been fitted to the carbonyl positions for some residues. A poly glycine chain is fitted in both directions to the map, and then unrestrained density fitting is carried out. The ratio between the geometric error statistics for the two hypotheses is a senstive measure. If one direction is

highly favoured rather than the other, the trace is probably correct; if there is no clear preference the trace may well be wrong. It is particularly helpful in fitting Beta strands.

2.8 Torsion angle real space refinement

The application can carry out real space refinement against the torsion angles. Its major advantage is the reduction in the number of variables to fit the model to the density. Its major problem when used with proteins is that the backbone atoms form a long chain of highly correlated phi/psi/omega torsion angles. If no consideration is given to this correlation, then any zone greater than 5 residues cannot be refined because the torsions in this region will not vary during the refinement process. Diamond [7], used a block refinement procedure where the protein was broken into small zones of residues but this caused problems at the link points.

In X-AUTOFIT the correlation is broken by running alternate cycles of refinement with and without restraints between the C=O and N atoms of a peptide link. This allows single residues to move independently of the rest of the chain at some stages during the refinement process. This procedure has a high radius of convergence and has been found to be a very successful method to refine the polymer chain. Another advantage of using torsion angle refinement is that the variables refined are almost independent to the traditional xyz(B) parameters used for X-ray refinement.

As the refinement is carried out in real space, phase information contributes to the fitting. Although this can provide more information to help fit the variables, it can also result in bias towards existing errors. Hence this algorithm can only be used as a tool for aiding model building rather than as a general method for protein refinement.

2.9 Monte Carlo torsion search methods for map fitting

The most difficult part of model building a protein is to fit the loops and termini. There are many reasons why this is so. Firstly there is not a simple set of secondary structure rules to restricts the possibilites for model building. Secondly as the termini and loops are commonly located on the surface of the molecule, they are very mobile and often poorly defiuned in the electron density. Thirdly it is possible that these regions were truncated by the mask during the solvent flattening process. Given these considerations, any automated method that could model build loops and termini would significantly speed up the map fitting process.

The method uses a novel algorithm that searches a series of backbone conformations for loops and termini, and chooses a good fit as one which has overlap with density, and good van der Waal contacts. The electron density near the site of the loop is first masked so that any part which overlaps closely with coordinates within the molecule, or in a symmetry equivalent position is excluded. A random number generator produces random backbone conformations for the required peptide length. These random conformations are then checked to see if they can approximately bridge the gap in the model and rejected if of the wrong length. (This test cannot be done for terminal fitting.) If this test is passed the overlap with density is calculated and scored. The top ten are displayed, colour coded by fit. The range of density fits is also displayed. The search can be stopped when a reasonable solution is observed or when the multiple solutions converge towards a single conformation.

The algorithm checks approximately 2500 conformations each second but the search method adds 2 more torsions to search fully (phi and psi) for each new residue and the procedure soon becomes impossibly slow. Even though the algorithm has a theoretical maximum of 25 residues, calculations of this size are not feasible. Nevertheless this search method has been used successfully on segments of up to 6 residues taking times in the vicinity of tens of minutes.

2.10 Regularisation

X-BUILD can regularise a structure using line minimisation of the derivatives. It reduces the deviation from ideality for the following: bonds, angles, chiral centers, planes, torsions, and non-bonded contacts to atoms within the molecule, and to symmetry equivalent molecules. Usually the crystallographer will use this to tidy the structure, but it is possible to turn on some defined torsions and non-bonds for various procedures. Some atoms can also be left fixed.

The new algorithm was developed to be used in interactive editing. As well as regularising a zone after building is complete, it is possible to define a zone of residues to be continually regularised. Individual atoms within this zone may be moved, while the regulariser attempts to preserve good overall geometry. This tool is extremely powerful when making large changes to protein structure, and has been found to be invaluable when rebuilding after X-ray refinement.

The algorithm is fast enough to allow the interactive editing of a zone of up to 8-10 residues on an R4000 base level indigo, and up to 50 residues on an

Extreme/High impact machine. This is probably more than would normally be edited during a model building process.

2.11 Other algorithms implemented as part of X-AUTOFIT

Several general functions have been written specifically for use in this application. They have been optimised to make the process of map tracing and model building as easy as possible.
The non-bonded and symmetry equivalent contacts use a lattice search algorithm which is currently the most efficient method for finding interactions between 3D points in space. Atoms can be bonded at a rate of 10,000 per second which generally means that there is no delay in updating the whole display over the local working region.

A function to generate rotational conformers around a bond has been written that can generate 10,000 of these per second per atom. This is obviously important for loop fitting and the real space refinement methods, but also means that any manual driven rotations around bonds appears smooth. The function that determines the overlap between coordinates and map uses both linear interpolation and quadratic interpolation as the former is quicker to use when far from a minimum, and does not result in false minima, but the latter results in better solutions when close to a minimum. This overlap function has also been optimised to allow refinement to proceed at a rate associated with normal xyzB reciprocal space refinement. Most of the electron density skeleton analysis uses a tree search function which is particularly applicable to this type of 3D network. This means that C_α placement, main chain/side chain bones assignment, and map quality which all require rapid pathway analysis of the bones can be accomplished instantaneously.

A novel fast contouring routine has been written for map masks which can contour 1.5 million points per second. This makes interactive editing of the mask a realistic possibility even on moderately sized machines.

3 The use of X-AUTOFIT and X-BUILD for real problems

All the tools to be described are controlled by a series of 12 palettes that contain up to 30 tool bars within the program QUANTA. Each palette is dedicated to some functionality such as map masking or sequence assignment. Each tool bar activates a facility or command in a modal form. It is possible to open all 12 palettes at one time although this is likely to obscure any other information on the screen. Therefore most crystallographic processes can be accomplished using only 2 palettes, and these can be opened, used, and closed at any time during the use of X-AUTOFIT and X-BUILD. Although X-AUTOFIT and X-BUILD represent different functionalities, they are completely integrated. For example, functionalities of the two modules can be combined. so that a map mask could be used as a limiting region for the C_α tracing process to prevent the building into symmetry or bones used with an omit map to build a difficult loop.

3.1 Electron density skeletonisation - bones

The electron density skeleton is controlled by a single palette with 12 options. The user can turn on and off bones at any time within X-AUTOFIT, and these are automatically generated to cover any part of displayed map, and updated when the screen origin is moved.

When editing bones for uses as a map mask it is useful to delete large parts of the bones that represent symmetry overlap. It is possible to delete all the connect bones from the selected bones point on the screen with a single action. This uses the tree search function to determine the connected bones, and hence is essentially intantaneous. It is possible to delete all the unwanted regions of bones to determine a map mask in the order of minutes. It is also possible to delete single branches of bones, change bones classification, or undo the last edit of the bones.
A tool calculates the symmetry related bones, usually in under a second to give a reduced representation of any symmetry related bones, including non-crsytallographically related bones. The reduced representation is used as these require less vectors to draw, and are therefore easier to manipulate; and since the normal use of bones is to only see where they overlap the "real" bones, then no information is lost by using a reduced representation.

The bones in X-AUTOFIT can therefore be used to determine a map mask, where the crystallographer would normally calculated a large volume of bones, and edit these until there are no symmetry overlaps. The map mask is then calculated to cover all the bones present in the volme described by the current active volume of map. The other use of X-AUTOFIT bones is to work with a small volume an use this as a guide in the C_α-tracing of

the electron density. For sphere is 12A the bones can be recalculated in around 1 second so that the crystallographer can experiment with the bones parameters, and move around the map with very little overhead associated with recalculation of the skeleton This procedure is described the next section.

3.2 Building chains of C_α atoms.

The C_α-tracing palette is designed for the initial interpretation of electron density calculated using the phasing methods of SIR/MIR and MAD. At this stage there is large error in the phase information, and the maps are on the edge of interpretability. It has already been shown that the use of bones can greatly aid in the intpretation of electron density [1] because the use of "lines" instead of solid regions dilimited by a net gives a many more visual views to the crystallographer experienced with stick model representations of proteins. The use of bones is therefore integral to the use of the C_α- tracing in X-AUTOFIT. There are two basic methods of building into the electron density a C_α-trace, the first is the placement of secondary structure elements into recognisable pieces of map (ie helices and beta strands) followed by rigid body refinement, and the second is to use the semi-automated C_α placement algorithm.

Recognisable secondary structure can be fitted quickly using sections of strand and helix, and then the fit of these to density can be improved with the rigid body refinement. These fitted sections can then be extended with the semi automated C_α fitting. To extend the chain trace, a tool is provide to fit the next C_α atom using the rule based fitting algorithm previously described. The crystallographer can adjust the postion of the atom by; picking a point on the bones, to which the C_α-C_α bond is pointed at, picking a conformation from the C_α geometry surface, or adjusting the open angle and torsion angle using the dials. A tool to show all the bones points 3.8Å from the previous C_α atom indicates all the possible paths the C_α trace can take, and offers a useful visual clue for building.

The application allows any number of fragments to be built, deleted and merged during the building process allowing a great degree of freedom for the crystallographer. It is also possible to select any fragment as active at any time using a point and click procedure. This action makes the nearest C_α atom active. If this C_α atom is at a terminal of a fragment then the user can extend the chain from this position, otherwise the user can change the x/y/z coordinate of this atom using the dials. The build direction is defined by the current active C_α atom, so the chain trace can be extened in either direction using the same tools and information. The polarity of the chain can be defined at a later stage using sequence alignement information and the auomated C_α direction detection tool.

The next stage of the C_α-tracing is probably the sequence assignment, which is described in the next session, and can be carried out on a C_α fragment of any size, although larger fragments are usually easier to determine.

3.3 Sequence assignment

The sequence palette controls the fuzzy alignment algorithm, and has two sub-palettes that list the 20 common amino acids, 10 woolly descriptions of residues and unknown. A tool allows the reading of a sequence as a one letter sequence file. After reading in a sequence the one letter code is displayed in lower case at the top of the model window. All segments built are then checked to see if any C_α has any residue description that is not unkown, and this is aligned. If a unique sequence match is determined then the sequence position that this represents is changed to upper case on the one letter sequence. The current active fragment is then aligned and a match of any kind is shown using the arrow indicators.

To change any alignment, or to start from scratch, the crystallographer needs only select the current active C_α atom, and then assign a residue type from the 31 options allowed. The program instantly shows any weighted match on the sequence table in both a forward and backward direction using arrows. The weight of the fit to the sequence based on the fuzzy descriptions of the residues types is show by the thickness of each arrow. If a unique sequence is determined, then the the sequence table is upper cased for this region.

Tools on the sequence palette relate to the current active segment. If a unique sequence has been determined for a fragment, but the crystallographer is not sure whether this is correct, then the alignment can be assigned NOT unique. This means that the sequence table in this region is available for alignment by other fragments. The uniqueness of a fragment can be set back with the same tool. If the program determines a unique hit for a fragment when a residue is changed to a deciding type, then the program will update all the C_α

atom types respectively. If, on inspection, this is found to be incorrect then the last fuzzy alignment can be returned. An alignment can also be cleared if deemed to be entirely wrong.

3.4 Generating an all atom model

An automated procedure is provided to generate an all atom model from the C_α-trace to use in X-build general model building part of the application. This tool generates an all atom model using the assigned sequence by real space refinement at about 5 residues each second on an R4000 SGI workstation. If no sequence information is specified a polyalanine trace is generated.

3.5 Model building

There are two palettes for general model building of macromolecules and two further sub-palettes for editing and colouring. All the tools that are normally associated with manual editing of proteins are provided, plus some powerful semi-automated modelling facilities with novel interfaces. The manul edit tools include, move atom, move zone of residues, change chi angles, change peptide plane, and edit the phi and psi angles of the first or last four residues of a segment.

The side chain of amino acids can be mutated to one of the 20 common amino acids with a single tool button, and after replacing the side chain the application fits the atoms to electron density by real space refinement. Geometric conformations for amino acids can be set to one of the the Ponders and Richardson likely conformers, or to a Sutcliff conformer. In each case the program shows all the non- bond interactions and the energy of interaction with the surrounding atoms.

The main power of the model building facility in X-BUILD results from the real space refinement algorithms and the interactive regulariser. There are tools to fit main chain atoms by real space fitting, fit side chain atoms by real space fitting, and also a tool that allows the movement of the C_α atom whilst real space fitting the side chain of that residue. The latter case is used when there is error in the position of the C_α position (or the chiral volume), and the real space fitting of the sidechain alone does not produce a sensible solution. A tool also allows a single residue to be refined by real space torsional angle refinement. This can also be used on amino acids, ligands and water atoms.

A tool allows simple regularisation of a zone of residues, the editing of atom positions while a zone of residues is actively regularised, and editing a single residue while it is regularised. All of these tools support disulphide bridges, optional non-bond interactions, and optional phi-psi restraints. The use of phi-psi restraints during geomtry refinement is open to abuse because it is possible for a crystallographer to set up the application to refine all the phi-psi values to the nearest allowed region on the Ramachandran plot to "improve" the geomtry at the detriment to the quality of the structure. This facility must therefore be used with caution. Its principle use is for modelling with low resolution experimental data and where there is no experimental data for general modelling of proteins. Fixed atoms are supported for all the tools for regularising coordinates, and the application automatically sets up fixed atoms where the regularise zone joins the rest of the protein structure by covalent links.

There are two futher palettes that are useful for model building. The first is to access the add/delete palette, and the second is to change the colouring of the active molecule quickly. The add delete palette allows peptide links to be made and broken to change the connectivity of the polypeptide chain, and allow terminal editing in the middle of a chain. Tools are available to delete residues, and ranges of residues, and also add residues at the terminus of a segment. The application carries out a tree search of conformations about the last phi and psi angles to fit the added residue to density, and then fits the side chain. This is a quick way of extending a terminal as density appear during refinement, as the application automatically adds the new residues to density.

X-AUTOFIT also has a facility to deal with partially ordered regions of the structure. Alternate conformations can be added and edited with any of the commands found elsewhere in the build palette. The application supports two types of alternate conformations. In the first case, where only the side chain is disorder, any changes automatically keep the B-conformer main chain atoms superimposed on the A-conformer positions. The second mode allows free editing of both the A and B conformer, so separate loops can be generated with the two different main chain conformations. The add/delete palette allows the changing of the C- terminal oxygens to required types, and also has a simple tool for renumbering the sequence ID's for each segment.

The colour palette allows the atom colour to be quickly changed to some simple colour scheme; by element type, alternate conformations different, by build progress, by fit to density, by temperature factor, and by occupancy.

A second model building palette contains tools for real space torsion angle refinement, rigid body refinement,

Monte Carlo loop fitting, Monte Carlo terminal fitting and a "do all" tool. All of these work on regions of structure. The refinement and loop/terminal fitting algorithms have been described earlier and provide powerful automated methods for model building proteins. With the interface via a GUI it is possible to try these on a "see whether it works" basis, and save/reject the changes depending on the outcome of the results.

The do all tool allows, for example, all the waters to be refined in turn, and the application checks for each atom/residue: mean atomic deviation, maximum atomic deviation, non-bonding , and density fit on completion. If any of these checks are worse than the user defined limits then the application will stop and allow manual intervention. So for example, if during the automatic refinement of all water molecules, one refines to a bad position, where there is no reasonable density, the application will stop and allow the crystallographer to delete this rogue water. Subsequent use of the do all tool will default to starting at the next atom/residue in the structure. This tool is invaluable now that structures are becoming so big, and often many hundreds of water positions can be identified.

3.6 3D text notebook

X-AUTOFIT and X-BUILD support a 3D notebook facility that allows the addition of the text strings at any point in 3D space. This allows the placement of comments about problems encountered during the building for subsequent building sessions. The 3D text editor also has some predefined list of text information such as termini or ligands. It is possible to advance to each text string using a single tool, and the map, bones and screen centre is updated about this position. This tool allows a crystallographer to quickely navigate around a large protein structure.

3.7 Validation

X-BUILD can carry out validation of some simple geometrical properties. This checks that the chiral, pro-chiral, planes, and nomenclature rules defined in the protein databank submission requirements are correct. The validation procedure will also check that the hydrogen bonding of glutamine, asparagine and histidine are maximised. The validation errors are placed in the 3D text editor described in the previous session, so allow the use to look and correct the particular problem determined during the validation procedure.

4 Summary

The aim of this program is to provide tools for the crystallographer to reduce the time taken for each step of this process but still allowing the experience of the crystallographer to determine the structure. The bones calculation is rapid, and allows the crystallographer to experiment to give the best representation for each part of the map. The C_α building tools provide many hints to aid the process of generating the main chain of the protein. Since the real space refinement procedure requires only the C_α positions to generate the entire residue, a significant saving in time and effort occurs without detracting from the excitement of determining a new structure. The sequence alignment is an extremely powerful tool which allows the use of woolly descriptions of residue type when the map is poor. While the model building section allows the use of grid, gradient and Monte Carlo methods of torsion angle real space refinement to aid in the model building process. The procedure described here can make significant savings in time for model building, typically, from days to hours, and weeks to days.

This work was supported by Glaxo Group Research, Molecular Simulations Inc [8], and the SERC.

5 Acknowledgements

I would like to thank all the crystallographers in the Protein Structure Research Group at York for testing the software and suggesting improvements during the application develoment. In particular, Maria Turkenburg who used the application in the very early stages when the program was not particularly helpful. I would like to thank Eleanor Dodson and Kevin Cowtan for many constructive conversations about crystallography, and Rod Hubbard, Liz Potterton and Helen Kemish for their invaluable help with QUANTA.

References

[1] J. Greer, “Three-dimensional Pattern Recognition : An Approach to Automated Interpretation of Electron density Maps of Proteins” J.Mol. Biol, Vol 82, pp 279-301, 1974

[2] A.T. Jones, J.Y.Zou, S.W.Cowan, M. Kjeldgaard, “Improved methods for building protein models in electron density maps and the location of errors in these models“ Acta Cryst, Vol A47, pp 110-119 , 1991

[3] K.D. Cowtan and P.Main, “Improvements of macromolecular electron density maps by simultaneous application of real and reciprocal space constraints”, Acta Cryst, Vol D49 No 1, pp 148-157, 1993

[4] T.J. Oldfield - unpublished algorithm

[5] T.J. Oldfield and R.E. Hubbard “Analysis of C-alpha geometry in protein structures”, Proteins: Structure Function and Genetics, Vol 18, No 4, pp 324-337, 1994

[6] M.H.Moore, J.M.Gulbis, E.J.Dodson, B.Demple, P.C.E.Moody, “Crystal structure of a suicidal DNA-repair protein - The ada O-6-methylguanine-DNA methyltransferase from escherichia-coli”, EMBO Journal, Vol 13, No 7, pp 1495-1501 1994

[7] R.Diamond, “A real-space refinement procedure for proteins”, Acta Cryst Vol 27, pp 436-452, 1971

[8] MSI 9685 Scranton Road, San Diego, Ca 92121-3752, USA.