ELSEVIER

Improved coordinate reconstruction from stereo diagrams

Rob W. W. Hooft and Gerrit Vriend

EMBL, Heidelberg, Germany

A program has been written that reconstructs three-dimensional coordinates for a protein structure given a stereo C_α diagram. Initial three-dimensional coordinates are determined using an algorithm similar to the one used by Rossmann in the program STEREO. Thereafter the coordinates are refined such that the stereo image based on the reconstructed three-dimensional coordinates optimally fits the scanned stereo image while normal C_α stereochemistry is enforced.

Keywords: stereo image, structure deposition, coordinate reconstruction

INTRODUCTION

The tremendous increase in knowledge about protein structures over the last few years allows experimental scientists using molecular modeling techniques to gain a deeper understanding of the molecules they work with. Consequently they have a better ability to predict how to design new characteristics by modifying existing proteins. A basic requirement, of course, is the availability of three-dimensional coordinates. Crystallographers are solving protein crystal structures at an ever-increasing rate, and most coordinates are, fortunately, sent to the Brookhaven Protein Data Bank repository (PDB).[1] However, regularly coordinates are not made available or become available only long after the initial publication describing the structure. This delay can either be due to commercial interest, or simply because the authors want to extract more information before other people can do so. The final result is that no coordinates are available for a large number of interesting molecules for which the structure was published.

These problems were addressed many years ago by M. G. Rossmann. To solve the problem Rossmann wrote a computer program that could regenerate three-dimensional coordinates from the information contained in C_α stereo diagrams. This program, called STEREO, is distributed by the PDB. Coordinate reconstruction requires that the two-dimensional positions of the C_α coordinates in both pictures are measured (either in a half-automatic manner, or simply by measuring them with a ruler). These coordinates, optionally supplemented with additional information about the secondary structure, are input to a least-squares procedure that refines three-dimensional coordinates and the stereo angle used to generate the picture. Many programs have been written on the basis of Rossmann's principles. In most cases the STEREO source is incorporated in a user-friendly, digitalization program. The final reconstructed three-dimensional C_α coordinates normally must be processed by a dedicated program to complete the model.

Oldfield and Hubbard[2] have written an elegant program (called EXTRACT) that fully interactively reconstructs three-dimensional coordinates. Their method centers around the display of a digitized image of the diagram overlaid on the reconstructed image on a graphics workstation with stereo viewing capabilities.

Here we describe a reconstruction method that combines some of the principles used by STEREO on the one hand, and EXTRACT on the other. The addition of a fully automatic optimization, however, reduces the amount of manual labor to a minimum. For the STEREO approach the accuracy of the initial 2D coordinates determines the accuracy of the result, so that much time is spent digitizing. In the EXTRACT approach, the human interactively positions all C_α atoms while the program imposes geometrical constraints. In our approach, a relatively rough input of 2D coordinates suffices, because the program can optimize them through comparison of a regenerated stereo line drawing with the scanned image. Automatic correction is possible as long as no residues are skipped in the initial 2D coordinate input. Through this approach, the final quality of the reconstruction depends only marginally on the precision of the initial tracing.

METHOD

Digitalization

The first step of the reconstruction process is the digitalization and tracing of the stereo image. We use the scanning

Address reprint requests to: G. Vriend, EMBL, Meyerhoffstrasse 1, D-69117 Heidelberg, Germany.

Received 30 April 1996; accepted 30 May 1996.

Journal of Molecular Graphics 14:168–172, 1996

655 Avenue of the Americas, New York, NY 10010

S0263-7855/96/$15.00
PII S0263-7855(96)00055-0

facility of a Kodak (Rochester, NY) color photocopier to digitize the original stereo diagram at a resolution of 200 or 400 dots per inch. This image is converted to the "portable graymap" format[3] and imported in the "xfig"[4] program. The drawing facilities of this program are used to index the C_α traces. Indexing simply implies that all C_α positions are picked in the order in which they occur in the molecule.

The digitized stereo image and the C_α tracing information are imported in the DIGIT option of the WHAT IF program.[5] The stereo image can be provided in the same "portable graymap" format; the C_α tracing information must be provided in the form of an xfig metafile.

C_α coordinate reconstruction

Given are two series of 2D coordinate pairs $(X_{1,k}, Y_{1,k})$, $(X_{2,k}, Y_{2,k})$, $1 \leq k \leq N$ and an (approximately) corresponding matrix of intensity values $\mathbf{A}$ with elements A_{ij}. Since the two images of the stereo plot are normally next to each other, $Y_{2,k} \approx Y_{1,k}$.

Generation of the initial three-dimensional coordinates From the two sets of coordinates the 2D difference vector $(\mathbf{D}_k)$ between corresponding points is calculated:

$$D_{x,k} = X_{2,k} - X_{1,k} \tag{1}$$

$$D_{y,k} = Y_{2,k} - Y_{1,k} \tag{2}$$

From these, the translation vector $(\mathbf{T})$ between the two plots follows:

$$T_x = \frac{\sum_{k=1}^{N} D_{x,k}}{N} \tag{3}$$

$$T_y = \frac{\sum_{k=1}^{N} D_{y,k}}{N} \tag{4}$$

With help of the unit vector $\mathbf{R} = \mathbf{T}|\mathbf{T}|$, the apparent translation a_k for each pair of 2D points is defined as (Figure 1):

$$a_k = (\mathbf{D}_k - \mathbf{T}) \cdot \mathbf{R} \tag{5}$$

By using the length of the projection on $\mathbf{R}$ instead of using the length of $(\mathbf{D}_k - \mathbf{T})$ directly, the influence of the picking error in Y is reduced. The first reconstruction of 3D coordinates can now be made using

$$x_k^0 = \frac{1}{2}(X_{1,k} + X_{2,k}) \tag{6}$$

$$y_k^0 = \frac{1}{2}(Y_{1,k} + Y_{2,k}) \tag{7}$$

$$z_k^0 = \frac{a_k}{2\tan(S/2)} \tag{8}$$

Here S is an estimate for the rotation angle used to create the stereo plot (usually around 6.0°).

Obviously, the 3D coordinates obtained in this way are on an arbitrary scale. To bring them to a normal protein scale, the average length L^0 of the $N - 1$ 3D vectors between adjacent points is calculated:

$$L^0 = < \delta^0 (k - 1,k) >_{2 \leq k \leq N} \tag{9}$$

with

$$\delta^0(k_1,k_2) = [(x_{k_1}^0 - x_{k_2}^0)^2 + (y_{k_1}^0 - y_{k_2}^0)^2 + (z_{k_1}^0 - z_{k_2}^0)^2]^{1/2} \tag{10}$$

In a further step, L is determined like L^0, excluding all distances $\delta^0(k - 1,k) < 0.67L^0$ and $\delta^0(k - 1,k) > 1.50\ L^0$ (these extreme values can occur where the interpretation of the 2D plot is difficult owing to overlap of lines). Finally the coordinates are brought up to this initial scale:

$$x_k^i = x_k^0 \frac{d_{12}}{L} \tag{11}$$

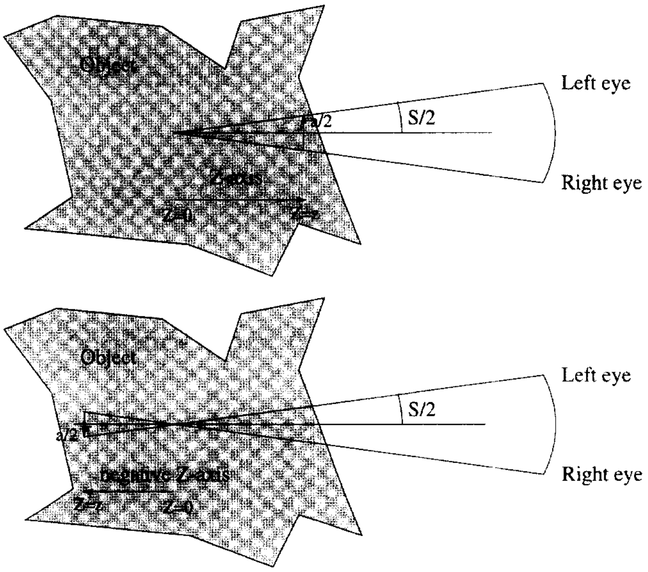

Figure 1. Initial Z coordinates can be obtained from the apparent translation a *using* $a_k/2 = z_k \tan S/2$ *[(Eqs. (5) and 8)]. Positive values of* a *result in positive values of* z *for points in the object toward the viewer, negative* a *result in negative* z *for points away from the viewer.*

$$y_k^i = y_k^0 \frac{d_{12}}{L} \tag{12}$$

$$z_k^i = z_k^0 \frac{d_{12}}{L} \tag{13}$$

with d_{12} the ideal distance between adjacent C_α atoms (3.80 Å).

To minimize the sensitivity of the algorithm to the stereo angle S, a least-squares optimization of the scale factors s_x, s_y, s_z in the three directions x, y, and z is performed, minimizing the penalty function:

$$\begin{aligned} P(s_x, s_y, s_z) = \sum_{k=2}^{N} \Big\{ d_{12} - \Big[& s_x^2(x_k^i - x_{k-1}^i)^2 \\ & + s_y^2(y_k^i - y_{k-1}^i)^2 \\ & + s_z^2(z_k^i - z_{k-1}^i)^2 \Big]^{1/2} \Big\}^2 \end{aligned} \tag{14}$$

Again, to prevent instabilities, only pairs for which $0.67L^0 \leqslant \delta^0(k-1,k) \leqslant 1.50\ L^0$ are taken into account during this scaling procedure. Scale factors are normally $0.8 < s_x, s_y, s_z < 1.2$. The final unoptimized coordinates are

$$x_k = s_x x_k^i \tag{15}$$

$$y_k = s_y y_k^i \tag{16}$$

$$z_k = s_z z_k^i \tag{17}$$

All transformations described so far are multiplied together into the 2D → 3D transformation $\hat{\mathbf{U}}$:

$$\begin{pmatrix} X_{1,k} \\ Y_{1,k} \\ X_{2,k} \\ Y_{2,k} \end{pmatrix} \begin{pmatrix} x_k \\ y_k \\ z_k \end{pmatrix} = \hat{\mathbf{U}} \tag{18}$$

Optimization: Target function The target function E for the optimization consists of two parts:

- The goodness of fit E_p between the regenerated stereo plot and the scanned plot **A.**
- A measure E_g of the violation of a number of geometrical constraints on the 3D C_α coordinates.

The geometrical constraints contained in our current program consist of two quadratic terms; the 1–2 distance violation term E_a and the 1–3 distance violation term E_b:

$$E_g = E_a + E_b \tag{19}$$

$$E_a = \sum_{k=2}^{N} [\delta(k-1,k) - d_{12}]^2 \tag{20}$$

$$E_b = \sum_{k=3}^{N} V_k \tag{21}$$

where the 1–3 violation V_k is defined as

$$V_k = \begin{cases} 0 & \text{if } \delta(k-2,k) > d_{13} \\ [\delta(k-2,k) - d_{13}]^2 & \text{otherwise} \end{cases} \tag{22}$$

In these equations $\delta(k_1, k_2)$ is defined similar to δ^0 [Eq. (10)]. The minimum 1–3 distance value d_{13} = 5.20 Å and the optimum 1–2 distance d_{12} = 3.80 Å were obtained from a study of 285 accurately determined protein structures from the PDB.

The goodness of fit for the stereo plot (E_p) uses the inverse transformation $\hat{\mathbf{U}}^{-1}$ of $\hat{\mathbf{U}}$ defined in Eq. (18). For each pair of adjacent 3D coordinates k and $k-1$, two lines (one for the left-eye and one for the right-eye view) are drawn through the intensity matrix **A,** and the value of all the encountered points is summed:

$$E_{1,k} = \sum_{ij \text{ in path 1, between } k \text{ and } k-1} A_{ij} \tag{23}$$

$$E_{2,k} = \sum_{ij \text{ in path 2, between } k \text{ and } k-1} A_{ij} \tag{24}$$

$$E_{p,k} = (E_{1,k} + E_{2,k}) \frac{\min(E_{1,k}, E_{2,k})}{\max(E_{1,k}, E_{2,k})} \tag{25}$$

$$E_p = \sum_{k=2}^{N} E_{p,k} \tag{26}$$

Equation (25) was chosen in this form to prevent a strong increase in one of the two plots from losing the track in the second one: normally the two equivalent paths in the two different images score approximately the same.

Optimization: Parameters The following parameters are optimized:

- The x, y, and z parameters of all points except the first and the last one.
- The z coordinates of the first and last points. Since locating these points in the digitized image is difficult (they are not defined by the intersection of two lines) the x and y coordinates are not refined.
- Depending on the progress of the optimization protocol, only a Z-scale s_z can be refined, or two scales s_{xy} and s_z, or three scale factors s_x, s_y, s_z.

Not all parameters must be refined at the same time: for a medium-sized protein this would require simultaneously optimizing 1 000 parameters. Considering that the coordinates of residues far apart in sequence are largely uncorrelated, the optimization is instead carried out in groups of three or more residues at a time. Scale parameters are refined in all blocks.

Thus, in a first optimization block positional coordinates of residues 1, 2, and 3 are refined, in the second block residues 2, 3, and 4, and this is repeated until the end of the sequence is reached. One cycle of refining all blocks is referred to as one "optimization round." Increasing the block size can change the results of the optimization, but the difference is marginal in comparison with the total reconstruction error.

The "Powell" method[6] is used for all optimizations.

Optimization protocol Since there are correlations between the parameters that must be optimized and the starting position might not be well defined, a straightforward maximization of the combined target function $E = E_p - wE_g$ would not lead to the best obtainable coordinates. A more careful approach is therefore followed.

Before any optimization, a conservative initial weight $w = w_i$ for the 3D C_α–C_α is chosen, to ensure that distance violations cannot disrupt the tracing of the stereo image:

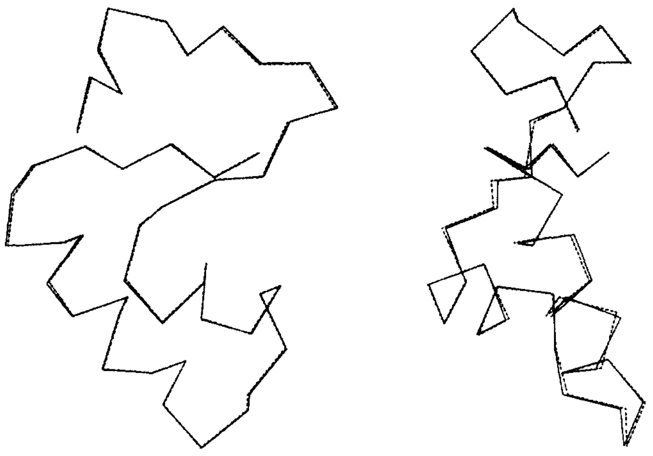

Figure 2. Superposition of real crambin C_α coordinates (dashed line) and coordinates reconstructed from a stereo imag (continuous line). The left image is an X–Y projection, the right image a Z–Y projection.

$$w_i = 0.01\,\frac{E_p}{E_g} \qquad (27)$$

At this early stage, the calculated line drawing uses a line width of three pixels, such that inaccurately traced lines can be found. Also at this stage, none of the three scale factors is refined. During a few rounds of optimization, the line width is reduced stepwise. When the line width is reduced to its minimum of 1 pixel, the lines are anti-aliased to allow for subpixel accuracy in the coordinate determination.

When convergence is reached, a scale factor s_z for all coordinates is introduced. After reoptimization a combined scale factor s_{xy} is introduced. After yet another round of optimization scale factors s_x and s_y replace the combined s_{xy}.

Finally, to improve the geometry of the molecule, the value of w is repeatedly doubled, optimizing the target function at every step. This is done until all distances between adjacent C_α–C_α are within 0.01 Å of the ideal distance d_{12}.

Reconstruction of the complete coordinate set

Several methods have been described to reconstruct a complete coordinate set starting with only C_α coordinates (for an overview see, e.g., Holm and Sander[7]). A full description of our method is beyond the scope of this article. Briefly summarized, a database search for fragments of five residues with similar relative C_α orientations is performed as described by Jones and Thirup.[8] The backbone coordinates of these fragments are used to reconstruct the backbone of the model. Second, position-specific rotamers are used to complement the model as described by Chinea et al.[9]

This reconstruction of the full atomic coordinate set from the C_α coordinates functions very well, but is not error free. Fortunately, one often has more information available than only the C_α stereo diagram. In most articles a stereo plot of the active site is given. In such cases we manually place the active site residues and use only the position-specific rotamer search method for the rest of the molecule.

EXAMPLE

A stereo plot of a C_α trace of crambin (Teeter,[10] PDB file 1CRN) was produced, and scanned as a grayscale image at 200 dpi. The scanned image was digitized, and 3D coordinates were reconstructed using the protocol described above.

Figure 2 shows two plots of the reconstructed C_α coordinates superimposed on the original coordinates. The RMS error in the C_α coordinates is 0.23 Å. To disseminate this error we redid the optimization of the stereo image overlap starting from the real C_α coordinates rather than from roughly estimated coordinates. This resulted in an RMS error in the C_α coordinates of 0.09 Å. This indicates that in this case about half of the error arises from printing and scanning the image, and the other half results from the imperfections of the optimization method. This is confirmed by the decomposition of the RMS errors into the components along the axes: RMS_z = 0.204 Å is only three times larger than RMS_x = 0.078 Å and RMS_y = 0.068 Å. A comparison with the all-atom RMS after a full reconstruction (1.06 Å) shows that the final amino acid-building step is responsible for the largest part of the reconstruction error.

DISCUSSION

Three-dimensional coordinates often are of great value for experimental scientists. Even if the coordinates are not very accurate, one can normally draw many conclusions about potential binding sites, proteolytic digestion sites, groups (not) to be attached to ligands, etc. If coordinates are lacking but a stereo diagram is available, our program can aid the experimentalist by reconstructing the three-dimensional coordinates as precisely as possible. We routinely use this program to generate three-dimensional models that can be used by our colleagues who use coordinates to design experiments in many different molecular biology disciplines.

There are several obvious improvements that can be implemented if higher precision is required.

- Use of Z coordinate-dependent line thickness where applicable
- Use of multiple views (stereo and/or mono) when available
- Incorporation of reconstruction of side-chain and ligand coordinates if these are shown in the diagram

We have, however, good experience with the program in its present implementation. Tests indicate that by far the largest portion of the final error in the all-atom representation is caused by the amino acid reconstruction, not by errors in the C_α positions.

AVAILABILITY

The complete subset of WHAT IF that can perform digitalization is available as a free standalone program WHAT_DIGIT. The program can be downloaded from our ftp site **swift.embl-heidelberg.de.** The freely available packages "xfig" and "pbmplus" can also be found there.

ACKNOWLEDGMENTS

We thank Francis Stewart, Henk Stunnenberg, and Chris Sander for stimulating discussions. Hans Döbbeling and his team skillfully maintain our computer systems. We thank Karina Krmoian and Brigitte Altenberg for technical assistance.

REFERENCES

1 Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer E.F., Jr, Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. The protein data bank: A computer-based archival file for macro-molecular structures. *J. Mol. Biol.* 1977, **112,** 535–542

2 Oldfield, T.J. and Hubbard, R.E. EXTRACT: A program to extract three-dimensional coordinates from stereo diagrams of proteins. *J. Mol. Graphics* 1995, **13,** 18–23

3 Poskanzer, J. Extended portable bitmap toolkit: pbmplus. 1991 The program suite `pbmplns` can be obtained in anonymous ftp from `ftp.x.org` in directory `/R5`-contrib

4 Smith, B.V. Xfig—Facility for interactive generation of figures under X11. 1995. The program `xfig` can be obtained via anonymous ftp from `ftp.x.org` in directory `/contrib/applications/drawing-tools/xfig`

5 Vriend, G. WHAT IF: A molecular modelling and drug design program. *J. Mol. Graphics* 1990, **8,** 52–56

6 Press, W.H., Flannery, B.P., Teukolsky, S.A., and Vetterling, W.T. *Numerical Recipes: The Art of Scientific Computing.* Cambridge University Press, Cambridge, 1986

7 Holm, L. and Sander, C. Database algorithm for generating protein backbone and sidechain co-ordinates from a C-α trace: Application to model building and detection of co-ordinate errors. *J. Mol. Biol.* 1991, **218,** 183–194

8 Jones, T.A. and Thirup, S. Using known substructures in protein model building and crystallography. *EMBO J.* 1986, **5,** 819–822

9 Chinea, G., Padron, G., Hooft, R.W.W., Sander, C., and Vriend, G. The use of position specific rotamers in model building by homology. *Proteins* 1995, **23,** 415–421

10 Teeter, M.M. Water structure of a hydrophobic protein at atomic resolution. Pentagon rings of water molecules in crystals of crambin. *Proc. Natl. Acad. Sci. U.S.A.* 1984, **81,** 6014–6018