Acta Cryst. (1997). D53, 434–437

A Connected Set Algorithm for the Identification of Spatially Contiguous Regions in Crystallographic Envelopes

JOHN F. HUNT,[a]* FREDERIC M. D. VELLIEUX[b] AND JOHANN DEISENHOFER[a]

[a]*Howard Hughes Medical Institute and Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, TX 75235-9050, USA, and* [b]*LCCP, IBS J. P. Ebel CEA/CNRS, 41 Avenue des Martyrs, 38027 Grenoble CEDEX 1, France. E-mail: hunt@howie.swmed.edu*

(*Received 11 November 1996; accepted 27 January 1997*)

Abstract

A simple algorithm is described for the identification of spatially contiguous regions in crystallographic envelopes. In a single pass through the grid points of the envelope map, the occupied points are assigned to a series of locally contiguous sets based on consideration of the connections within single voxels. A spatially contiguous region is identified as the union of all of the locally contiguous sets that share an element in common. Therefore, chains of spatial connectivity are traced implicitly by performing simple set operations. This algorithm has been implemented in the program *CNCTDENV* as part of the *DEMON/ANGEL* suite of density-modification programs.

1. Introduction

Spatial connectivity of an appropriately contoured electron-density function is a fundamental property of the crystallographic maps of polymers at low and intermediate resolution. As a result, the presence of spatially contiguous regions has been exploited in a variety of phase-improvement procedures used in the crystal structure determination of polymers. Iterative skeletonization procedures rely on the snake-like connectivity of the polymer chain to attempt phase improvement at intermediate resolution (Baker, Bystroff, Fletterick & Agard, 1993; Swanson, 1994). The connectivity criterion can also be used to improve the quality of the molecular envelopes used in solvent flattening (Wang, 1985; Cowtan & Main, 1996) and non-crystallographic symmetry averaging procedures (Bricogne, 1976; Kleywegt & Jones, 1994*a*).

A molecular envelope is a binary map encoding the region of the unit cell occupied by the polymer as opposed to solvent; it is generally derived by applying an electron-density cut-off criterion to a map filtered to low resolution (Leslie, 1986). Large contiguous regions in such an envelope are likely to reside within the macromolecule, while small contiguous regions are likely to represent noise. Elimination of those small regions improves the quality of the envelope and generally yields better results in any density-modification procedure performed using the envelope (Kleywegt, 1994; Kleywegt & Jones, 1994*a*; Tête-Favier, Rondeau, Podjarny & Moras, 1993).

The identification of contiguous regions in three-dimensional space also has applications in theoretical and computational analyses of protein structure. For example, cavities in proteins can be viewed as spatially contiguous regions of unoccupied points after projection of the electron density of the protein atoms onto an appropriate three-dimensional grid (Delaney, 1992; Kleywegt & Jones, 1994*b*).

In summary, there are diverse applications for an efficient algorithm for the extraction of spatially contiguous regions from three-dimensional maps. We have developed such an algorithm based on the treatment of the contiguous regions as connected sets of points. The algorithm has been implemented in the computer program *CNCTDENV* which operates on crystallographic envelopes; this FORTRAN program is distributed as part of the *DEMON/ANGEL* density-modification suite (Vellieux, Hunt, Roy & Read, 1995).

2. Algorithm

The algorithm proceeds by assigning the occupied grid points to a series of locally contiguous sets, A_m, based on consideration of the connections between the eight points in each individual voxel in the map; these locally contiguous sets are called 'anchor sets'. Globally contiguous sets, C_l, are then constructed by combining all of the anchor sets that are mutually contiguous. Mathematically, this operation could be described as forming the union of all locally contiguous sets which posses an element in common,

$$C_l = \bigcup_{m_l \in L} A_{m_l};\ A_{n_l} \bigcap \left(\bigcup_{m_l \in L, m_l \neq n_l} A_{m_l} \right) \neq 0 \text{ for } \forall\ n_l \in L.$$

In this expression, L represents a set containing all indices m_l of anchor sets contributing to the connected set C_l. The advantage of this approach is that it requires only a single pass through the grid points of the map

because chains of spatial connectivity are followed implicitly in the process of forming a union of connected sets.

In practice, the envelope map is converted to an anchor-set map in the course of a sequential pass through all of the grid points $P(ijk)$. This process is shown schematically in two dimensions in Fig. 1. At every occupied point $P(ijk)$, the occupancy state is examined only for the seven other points in the voxel shown in Fig. 2 [*i.e.* the points with indices i or $(i-1)$, j or $(j-1)$, and k or $(k-1)$]. Because of the sequential nature of the assignment process, all of these points other than $P(ijk)$ have already been examined and assigned to one or more anchor sets, if they are occupied. The point $P(ijk)$ is assigned to the same anchor set if any other point in the voxel is occupied; otherwise it is assigned to a new anchor set. (Note that before any voxel can be treated, the occupied points on the three faces of the map with $i = 0$, $j = 0$, or $k = 0$ must be assigned to anchor sets using the two-dimensional procedure shown explicitly in Fig. 1.)

If the eight points in the voxel are assigned to different anchor sets, a spatial connection exists between these locally contiguous sets. In a formal treatment of the problem in terms of set theory, each shared element or grid point would be included in both of the connected anchor sets; however, there is no need to adhere to this procedure in a practical implementation of the algorithm. Instead, a record of the connection between anchor sets is stored in the form of a symmetric matrix with rows and columns representing the anchor sets. This matrix is called the binary connection matrix, **B**, because the elements assume a value of either 0 or 1 depending on whether the corresponding anchor sets are spatially connected (*i.e.* share an element in common). In the case that a connection is found between anchor

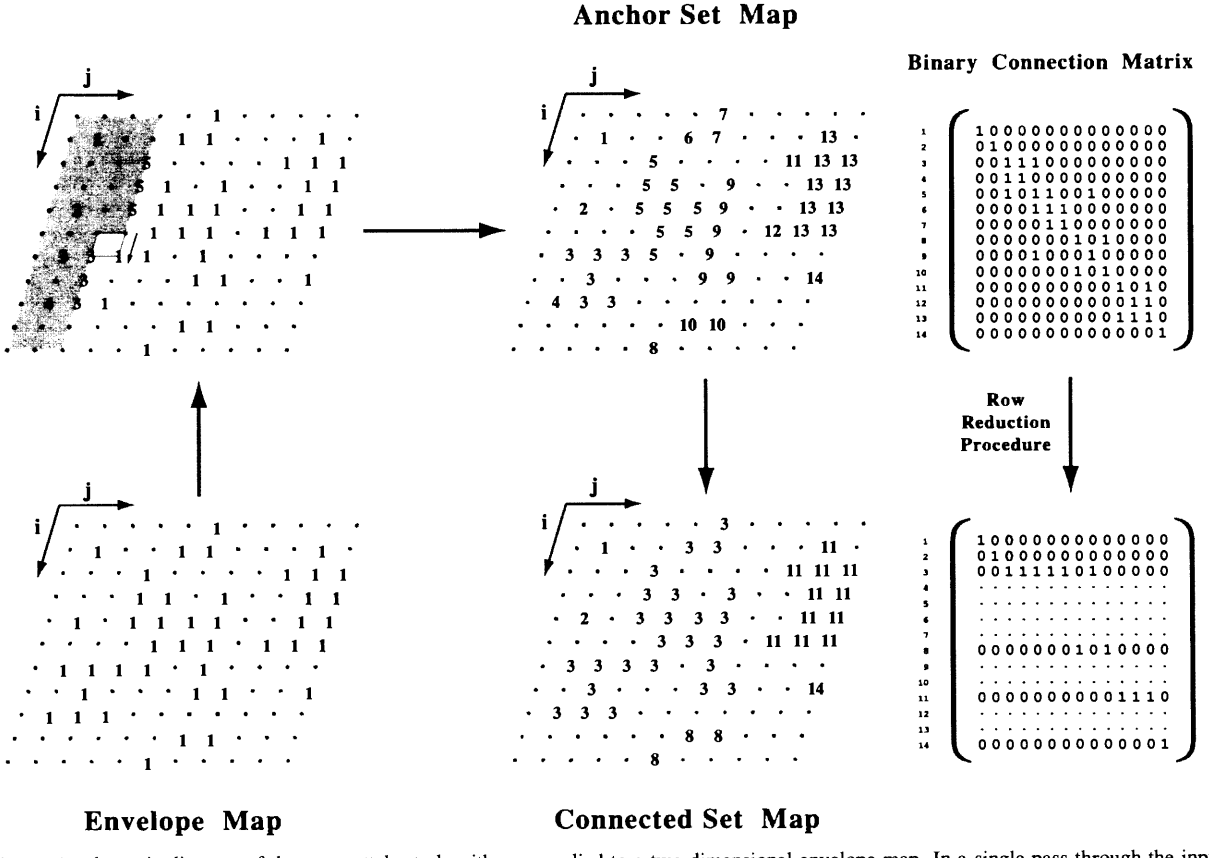

Fig. 1. A schematic diagram of the connected set algorithm as applied to a two-dimensional envelope map. In a single pass through the input map, the occupied points of the envelope are assigned to a series of locally connected anchor sets A_m. The numbers in the anchor-set map represent the indices m of the anchor sets to which the occupied points are assigned. The map on the upper left represents an intermediate step in the construction of the anchor-set map in which only the shaded area has been processed. The small arrow shows the direction of progression through the grid points of the map. The point about to be evaluated is at the lower right of the pixel outlined in the thin black box; the next step in the construction of the anchor-set map would be to assign this point to anchor set number 3 based on the local connectivity within the pixel. The observed local connections between pairs of anchor sets are recorded in the binary connection matrix **B**; row reduction (see text) of the binary connection matrix yields a series of connected sets C_l which can be encoded in the output map. (In the depiction of the reduced binary connection matrix, the rows that have been combined into rows of lower index are indicated by dots for the sake of schematic clarity.)

sets m and n, a value of 1 is assigned to the elements B_{mn} and B_{nm} of the matrix, which had previously been initialized to 0. Note that it is necessary to check for connections between the anchor sets in a given voxel even if the current point $P(ijk)$ is unoccupied; the reason for this requirement is illustrated by the connection between anchor sets 8 and 10 in Fig. 1, which would be missed otherwise.

Because of the sequential nature of the pass through the grid, the connections within three of the six walls of the current voxel have already been evaluated during consideration of previous grid points [*i.e.* the $(i-1)$ wall, the $(j-1)$ wall, and the $(k-1)$ wall]. In this context, the following three rules apply to the accounting of connections between the anchor sets assigned to the eight points of the current voxel. First, if there exist two or three edge connections, there cannot exist any unique diagonal connections within the faces or the body of the voxel. Second, if there exists a single edge connection, there can exist at most one unique diagonal connection which is on the face orthogonal to the edge connection and containing the point $P(ijk)$. And, finally, if there are no edge connections, there can exist at most one unique diagonal connection (*i.e.*, in this case, there cannot exist a unique body diagonal connection in addition to any single face diagonal connection).

It is possible to adjust the allowable spatial separation between points to be considered contiguous by changing the way diagonal connections are treated within a voxel. Usually, the body diagonals of the voxel are longer than the face diagonals which are longer than the edges. Thus, the 'diagonal mode' of the anchor-set assignment process can be adjusted to consider all three types of connections as valid or just a subset of them (*i.e.* only face diagonals and edges or exclusively edges). However, this adjustment gives only a small level of control over the strength of the connection required to consider a region to be contiguous, and more significant variations in this regard can be achieved by cyclical trimming/addition of surface points on the envelope (Delaney, 1992; Kleywegt, 1994).

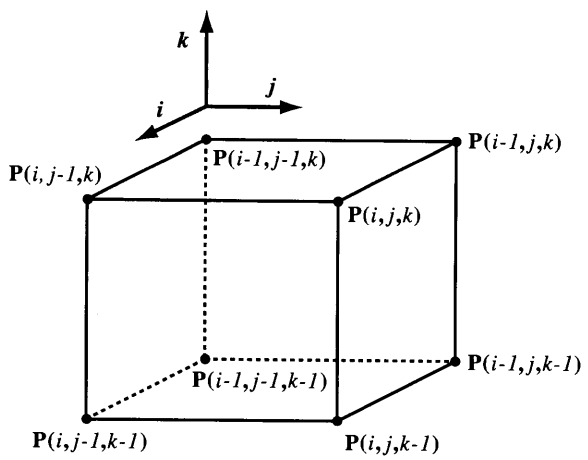

Fig. 2. An indexed voxel in the envelope map. The anchor-set map is constructed in a single sequential pass through the grid points of the map with index i varying the fastest and index k varying the slowest. Therefore, by the time point $P(ijk)$ is evaluated for assignment to an anchor set, the seven other points in the current voxel have previously been evaluated and assigned to anchor sets if occupied. The top face of the voxel represents the relevant geometry for a two-dimensional map as illustrated in Fig. 1.

The anchor sets can be combined into globally contiguous sets by applying a simple row-reduction procedure to the binary connection matrix **B** starting with the row of highest index and proceeding sequentially through all the rows of lower index. For each row m corresponding to anchor set A_m, the elements B_{mn} are scanned for a connection between this set and some anchor set A_n with a lower index $n < m$. If such a connection is detected, a bitwise logical OR operation is performed to combine the sets in the current row m with the sets in the connected row n, *i.e.* $B_{np} = (B_{np} \parallel B_{mp})$ for all columns p. For a given row m, this combination operation is performed only once, but it must be performed on the row corresponding to the connected anchor set A_n of highest index $n < m$ in order to ensure that other anchor sets with independent connections to A_n are systematically passed through to the final globally connected set. If no such connection is present, the row m is identified as a complete connected set because the current anchor set A_m can only have connections to anchor sets with higher indices that have already been combined into the current row in the binary connection matrix.

The size of every anchor set is recorded during construction of the anchor-set map in the form of a vector **S**, with elements S_m representing the number of envelope map points assigned to anchor set A_m. Therefore, the overall size of each globally contiguous set is readily evaluated as the sum of the relevant elements in this vector. In general, the largest connected set of points or the union of the top few largest sets yields the best envelope for use in non-crystallographic symmetry (NCS) averaging procedures, assuming that phases of reasonable quality are used to derive the initial envelope map. In the program *CNCTDENV*, specific connected sets or unions of connected sets can be designated for output into the new envelope file. The output procedure runs through the points of the anchor-set map and only sets them as occupied in the new envelope if the anchor set is an element of one of the specified connected sets.

3. Discussion

The program *CNCTDENV* processes an envelope with 150 grid points per edge in approximately 10 CPU seconds on a DEC Alpha AXP system running at 195 MHz. One advantage of this algorithm is that it does not make any assumption or approximation concerning

the pattern of spatial connections in the map. Therefore, all sets of points which are globally connected according to the operative local criterion are always identified and passed through to the output map after a single pass through the grid points of the input map. The current version of the program is designed for use in NCS averaging procedures and is ignorant of crystallographic symmetry. In order to apply the algorithm in solvent-flattening procedures, it would be desirable to have the program handle spatial connections which traverse the boundaries of the crystallographic asymmetric unit. A simple modification of the program would allow the identification of anchor sets that are connected by the application of crystallographic symmetry.

One limitation of the connected-set algorithm is that the number of anchor sets and, therefore, the size of the binary connection matrix is somewhat unpredictable. Given a grid with n_i, n_j and n_k points along its three edges, a pathologically disconnected map can be constructed with $(n_i n_j n_k/8)$ occupied points sharing no spatial connections (*e.g.* a map comprising 125 000 mutually disconnected anchor sets for a grid with 100 points per edge); in this case, the binary connection matrix would contain more than 10^{10} elements if a row were assigned for every anchor set. However, the size of the binary connection matrix can be significantly reduced by assigning a row in the matrix only for anchor sets with at least one connection to another anchor set (which reduces its size to zero in the case of the pathologically disconnected map). Furthermore, a modest reduction in the number of anchor sets can be achieved by checking the next voxel for previous anchor-set assignments in the case that none of the other points in the current voxel are occupied. Finally, conservation of the amount of computer memory required to store the binary connection matrix can be achieved by using a single bit to represent each element. In practice, the size of the matrix has not been a limitation in implementing the algorithm because, so far, we have observed experimental envelope maps to contain a total of at most 2000 anchor sets. While it is possible to implement the algorithm by storing a list of the pairwise connections between anchor sets without resorting to the use of a matrix, resolving the chains of connectivity is computationally inefficient because of the inherently indirect nature of the connections between many of the anchor sets as well as the redundancy of the connections between them.

One potential future application of this algorithm would be in a Bayesian solvent-flattening procedure that exploits the prior knowledge that the polymer regions of a molecular envelope should occur in large, spatially contiguous sets. Specifically, the connected-set algorithm produces a quantitation of the size of all of the spatially contiguous regions in the envelope. Bayes' rule could be used to estimate the probability that a spatially contiguous set of a given size occurs within the molecular boundary of the polymer based on some prior assumption about the distribution of the sizes of contiguous regions in the polymer *versus* solvent regions, *i.e.* some specific quantitative expression of the assumption that large spatially contiguous regions occur with high probability in the polymer region but with low probability in the solvent region. Additional prior assumptions concerning spatial connectivity could also be introduced by appropriate manipulations of the probabilistic envelope, for example by increasing the weight for points on the surface of the large, high-probability regions in the envelope.

The authors would like to thank Dr Hee-Won Park for productive conversations and suggestions, and Drs Lothar Esser and Mischa Machius for critical reviews of the manuscript. JFH was supported by a postdoctoral fellowship from the Jane Coffin Childs Memorial Fund for Medical Research. FMDV gratefully acknowledges financial support from the CEA and CNRS.

References

Baker, D., Bystroff, C., Fletterick, R. J. & Agard, D. A. (1993). *Acta Cryst.* D**49**, 429–439.
Bricogne, G. (1976). *Acta Cryst.* A**32**, 832–847.
Cowtan, K. D. & Main, P. (1996). *Acta Cryst.* D**52**, 43–48.
Delaney, J. S. (1992). *J. Mol. Graphics*, **10**, 174–177.
Kleywegt, G. J. (1994). *MAMA - The Manual*. University of Uppsala, Uppsala, Sweden.
Kleywegt, G. J. & Jones, T. A. (1994*a*). *Proceedings of the CCP4 Study Weekend: From First Map to Final Model*, pp. 59–66. Warrington: Daresbury Laboratory.
Kleywegt, G. J. & Jones, T. A. (1994*b*). *Acta Cryst.* D**50**, 178–185.
Leslie, A. G. W. (1986). *Acta Cryst.* A**43**, 134–136.
Swanson, S. M. (1994). *Acta Cryst.* D**50**, 695–708.
Tête-Favier, F., Rondeau, J.-M., Podjarny, A. & Moras, D. (1993). *Acta Cryst.* D**49**, 246–256.
Vellieux, F. M. D., Hunt, J. F., Roy, S. & Read, R. J. (1995). *J. Appl. Cryst.* **28**, 347–351.
Wang, B. C. (1985). *Methods Enzymol.* **115**, 90–112.