SUPPLEMENTARY MATERIAL

Details of the Procedure of the Average Distance Map Method

The method used mainly in the present work is described in Kikuchi *et al.* (1988) and let us give a brief survey of the method. We refer to this method as the Average Distance Map (ADM) method.

(1) Definition of Ranges by Separation Between Residues Along the Sequence of a Protein and the Calculations of Average Distances within Each Range

A range is defined as the length between two residues along a given sequence, i.e., a range is defined as M = 1 when $1 \leqq k \leqq 8$ where k = |i-j| and i and j are the residues numbers along the sequence. In the same way, $9 \leqq k \leqq 20$, $21 \leqq k \leqq 30$, $31 \leqq k \leqq 40$ and so on define respective ranges M = 2, 3, 4 ···. The average distances between Cα atoms of residues were calculated in each range using proteins with known structures.

A contact map is constructed by making a plot (i.e., defining a contact) on a map for a protein with unknown 3D structure, if the average distance of a pair of residues in a range defined above is less than a cutoff value determined by the method described in the following way.

(2) Definition of Cutoff Distances for Construction of ADM

A cutoff distance value for the construction of ADM of a given sequence is defined in each range so that the contact density of the whole real distance map (RDM) of the protein is reproduced (Kikuchi *et al.* 1988). The RDM for a contact map is constructed based on the actual 3D structure. In the present study, a contact on the RDM is defined as an inter-residue Cα atomic distance less than 15 Å. Regarding ρ_{av} as the average values of contact density of the entire region of a map, the value of ρ_{av} on the RDM can be roughly predicted by the formula, $\rho_{av} = \frac{C}{N}$ (Kikuchi *et al.* 1988), where N is the total number of residues and C is a adjustable constant (C = 36.12 in the present work) (Kikuchi *et al.* 1988). Cutoff distances for construction of an ADM of a protein are defined to reproduce a value of $\rho_{av} = \frac{C}{N}$. We define a different cutoff distance for a different range in the construction of ADM in contrast to the case of the construction of RDM where the same cutoff distance is used. Here, we make an assumption that the number of residue pairs that make a contact obeys the following Eq. (1) in a range M (Kikuchi *et al.* 1988).

$$P(M)_C = (\frac{D}{M})P(M)_t \quad (1)$$

Here, $P(M)_C$ is the number of amino acid pairs whose average distances in the range M is less than a cutoff distance, i.e., residue pairs to be plotted on ADM, and $P(M)_t$ is the total number of residue pairs in a given range, i.e., 210 pairs of residues minus the number of the pairs with statistically insufficient occurrence (Kikuchi *et al.* 1988). D is an adjustable parameter chosen so that the overall average density ρ_{av} of the ADM is close to the predicted value of ρ_{av} on RDM. Thus, we can construct a predicted contact map from only the sequence of a given protein based on the inter-residue average distances. The final map obtained is an average distance map or ADM. (In the construction of ADMs for azurin and titin, D = 1.25 and D = 1.40 are used respectively. We obtained the predicted $\rho_{av} = 0.286$ and 0.406 for azurin and titin respectively.)

(3) Calculations and Scan of Contact Density Difference

A contact density difference is defined as $\Delta\rho_i = \rho_i - \tilde{\rho}_i$ where ρ_i and $\tilde{\rho}_i$ denote the contact density of the triangle and trapezoidal parts, respectively, when the whole area of a map is divided into two parts by a line parallel to the abscissa at the i-th residue or by a line parallel to the ordinate at the i-th residue as illustrated in Fig. (**1A** and **1B**).

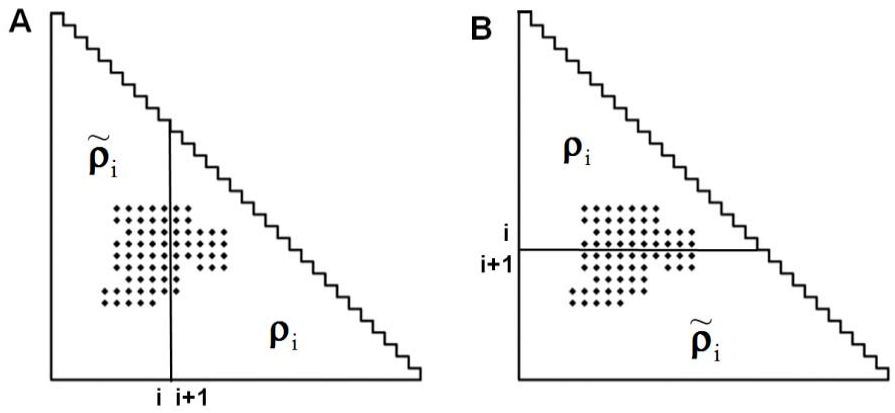

Fig. (1). A. Schematic drawing of a contact map divided by a line parallel to the ordinate at the residue i. An asterisk on the map denotes a contact. **B**. Schematic drawing of a contact map divided by a line parallel to the abscissa at the residue i. An asterisk on the map denotes a contact. The density of contacts is defined as $\tilde{\rho}_i$ for a trapezoidal part and as ρ_i for a triangle part.

This procedure is iterated from residues 1 to N, and then a scanning plot of contact density difference is obtained. We call the scanning plot produced by the division using the line parallel to the ordinate as horizontal scanning, and the plot produced by the line parallel to the abscissa as vertical scanning. h of $\Delta\rho_i^h$ and v of $\Delta\rho_i^v$ denote the horizontal and vertical divisions of a map respectively.

(4) Definition of Compact Regions

The maximum (peak) and minimum (valley) would be obtained in the scanning plot as the point of a large change of contact density values on a map. An example of a horizontal scanning plot of $\Delta\rho_i^h$ from 1 to N is shown in Fig. (**2**), and peak and valley appear at a and b in the figure at a large change of contact density values.

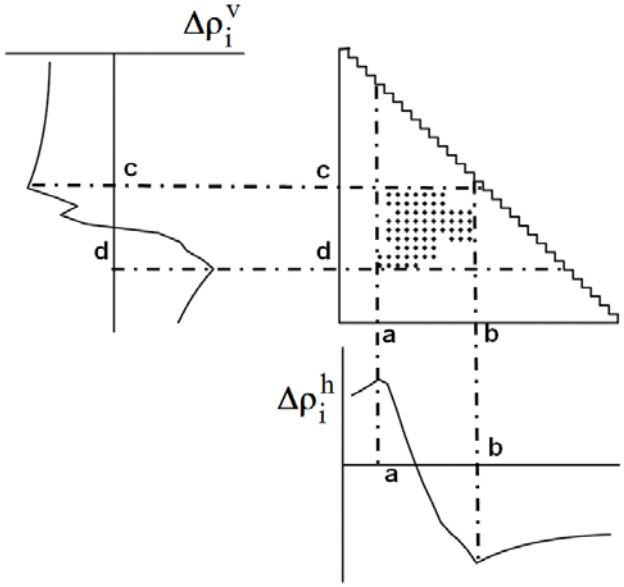

Fig. (2). A scanning plot of the contact density differences, $\Delta\rho_i = \rho_i - \tilde{\rho}_i$, from 1 to N on a map. When the scanning plot is produced by the division using a line parallel to the ordinate, we refer to this as horizontal scanning, i.e., a scanning plot is defined by the values $\Delta\rho_i^h = \rho_i^h - \tilde{\rho}_i^h$. In this figure, the scanning plots for a hypothetical contact map are drawn. In the horizontal scanning plot, we notice that a peak and a valley appear at a and b corresponding to a large change of contact density values. Likewise in a scanning plot defined by the plot of $\Delta\rho_i^v = \rho_i^v - \tilde{\rho}_i^v$, the differences in contact density defined by the division using a line parallel to the abscissa, is referred as a vertical scanning. In the vertical scanning plot, a peak and a valley appear at c and d, respectively. That is, peaks and valleys on a map appear at the boundaries of a high density contact area, i.e., a compact region.

The same situation is observed in the vertical scanning in the same figure. Thus, as noticed from this figure, the boundary of a compact region on a map can be detected by a peak and a valley appearing in horizontal and vertical scanning plots of contact density differences.

(6) Prediction of Location of Subdomains

A subdomain on a map can be defined by the positions of the peaks of scanning plots as shown in Fig. (**3**). This figure shows a hypothetical contact map with two compact areas near the diagonal.

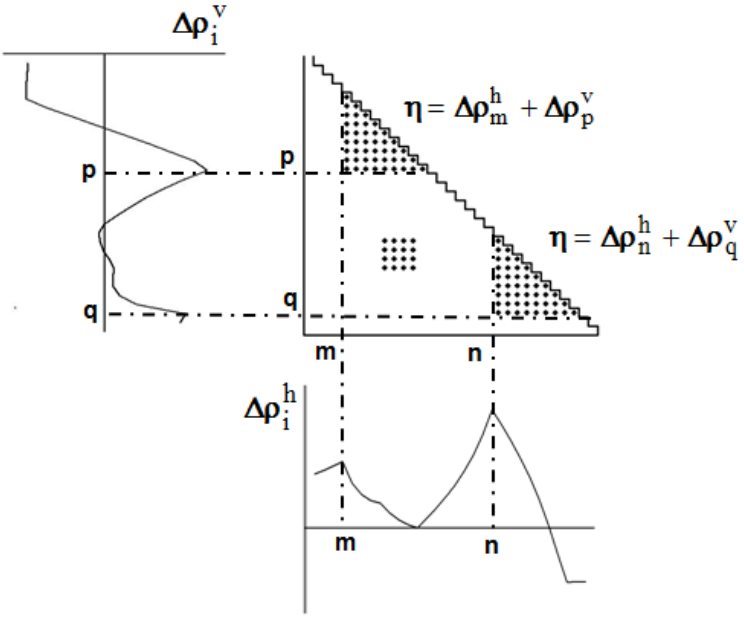

Fig. (3). A hypothetical contact map with two compact areas near the diagonal of a map with horizontal and vertical scanning plots. The peaks at residues m and n in the horizontal scanning plot and at residues p and q in the vertical scanning plot indicate the existence of two domains. The regions m-p and n-q on the map are the possible compact regions or domains in the protein. $\eta = \Delta\rho_m^h + \Delta\rho_p^v$ and $\eta = \Delta\rho_n^h + \Delta\rho_q^v$ denote propensities of the compactness of respective regions.

The horizontal and vertical scanning plots are also presented in the same figure. We recognize the existence of two domains by the peaks at residues m and n in the horizontal scanning plot and residues p and q in the vertical scanning plot. Thus, regions m-p and n-q on the map are predicted as possible compact regions or domains in the protein.

(7) Measure of the Compactness of a Compact Region Defined on ADM, η Value

The strength of the compactness of a region m–p can be measured by the η values defined by $\eta = \Delta\rho_m^h + \Delta\rho_q^v$ (Fig. **3**) (Kikuchi *et al.* 1988). Thus, based on this procedure, we can make a prediction on location of compact regions in a protein from only its amino acid sequence. The region with the highest η value can be defined as the maximum of a compact region. Other regions with high η values can be assigned as smaller compact regions (Kikuchi *et al.* 1988).

REFERENCE

T. Kikuchi, G. Némethy, and H.A. Scheraga, "Prediction of the location of structural domains in globular proteins". *J. Protein Chem.*, vol. 7, pp. 427-471, 1988.