The method based on singular value decomposition (SVD)

Ideal case

This analysis is based on the properties of the matrix representation of equivalence relations. Let us consider a relation $\Re$ defined as ‘belongs to the same domain as’. It is straightforward to show that this relation is reflexive, symmetric and transitive (assuming that domains do not overlap) and thus it is an equivalence relation. An equivalence relation can be represented by a binary matrix, *M*: if $i \Re j$ then $M(i,j) = 1$ else $M(i,j) = 0$. The matrix representation of equivalence relations is such that equivalence classes (i.e., domains in the case we are considering here) appear in this matrix as symmetric blocks with respect to the principal diagonal. (See Fig. 1b.)

When this matrix is factored using the singular value decomposition (SVD) [http://en.wikipedia.org/wiki/Singular_value_decomposition], we have

$$M = U * S * V^T, \qquad (12)$$

where S is an m-by-m diagonal matrix with positive or zero elements (the singular values) and U and V are m-by-m orthogonal matrices, the columns of which give the (left and right) singular vectors. This decomposition provides:

1. The number of domains (the number of non-zero elements in S, called the rank r of M);
2. The number of residues in each domain (the singular values);
3. The residues that constitute each domain (the non-zero entries in each column of U or V).

For instance, if we apply SVD to the *M*-matrix of 1atnA shown in Fig. 1b (corresponding to the following CATH domain definitions: *D1*= [4-34], [71-134], [337-372]; *D2* = [35-68]; *D3* = [136-181], [271-332]; *D4* = [182-267]), (see Fig. 1a), we obtain four singular values greater than zero: 131, 108, 86, and 34, all other singular values being zero. All the non-zero singular values are integers and correspond to the number of residues in *D1, D3, D4* and *D2,* respectively. Inspection of vector U_1 (or V_1^T) reveals that 131 elements, [4-34], [71-134] and [337-372], are different from zeros. Notice that these are the residues defining domain *D1*. A similar observation is made for the three other domains.

Observed N-matrix

The actual *N*-matrix contains information on the equivalence relation, but it is clearly not in the form of an *M*-matrix. The most obvious difference is that the former is not binary (see Fig. 1c). We therefore choose a threshold *T* and convert the *N*-matrix to the binary matrix *B* by setting $B(i,j)=1$ if $N(i,j)>T$, else $B(i,j)=0$. The situation is then analogous, but not completely similar, to the ideal case when we apply SVD to a binary *B*-matrix.

The number of domains is obtained by sorting the singular values from large to small and then counting them, not all the way down to zero, but down only to the first one for which the associated left and right singular vectors have the opposite sign: $U(k)=-V(k)$ where $U(k)$ and $V(k)$ are the *k*-th columns of *U* and *V* matrices, respectively.

The number of residues in each of these domains is determined as follows. Fig. 2 shows the sorted plot of the square of the elements of the singular vectors with the largest singular value for 1atnA with T value of 300 (Fig. 1f). For the ideal case of the equivalence relation matrix, we obtain a step function (the dashed pink line in the plots): all elements of the equivalence class (domain) have the same value; all others are equal to zero. The position of the vertical line coincides with the associated singular value, which gives the number of residues of the domain. For the observed case, we round the singular value to the nearest integer and then choose the nearby position on the plot (the green dotted line) for which the slope is maximal (the slope is infinite for the step function).

The residues belonging to a domain *a* are normally the first *nr(a)* of the sorted elements of the corresponding singular vector *U(a),* where *nr(a)* is the position of the maximum slope as determined above. After this step, all domains less than 40 residues are discarded and their residues are labelled as “idle”, i.e., they are not assigned to a domain. We apply a post-processing to assign these idle residues to the domains based on their distance to the domain centroids.

Applying the above procedure to the binary matrix shown in Fig. 1f, we obtain 142 singular values greater than 1, which are no longer integers. The first five have magnitudes: 128.1, 92.0, 69.7, 54.9 and 29.9, the rest of them decrease monotonically to zero. The analysis of the singular vectors and values provides four values 119, 85, 70 and 52 that correspond to the number of elements in the four domains and we obtain the following domains *X1* = [137-178], [273-349], *X2* = [5-23], [77-136], [356-360], *X3* = [189-258] and *X4* = [25-76], a relatively good agreement with the CATH domain definition.

Determination of the threshold value T

To each domain partition, obtained using a threshold value T, are assigned three quality measures, Q, Q_1, and Q_2. Q_1 measures the degree with which the input *B*-matrix deviates from the *M*-matrix computed from the derived domain definition,

$$Q_1 = \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} \left| B(i,j) - M(i,j) \right|. \qquad (13)$$

Q_2 is the fraction of “overlapping” residues,

$$Q_2 = \frac{1}{m} \sum_{a=1}^{nd} \sum_{b=a+1}^{nd} O(a,b), \qquad (14)$$

where *O(a,b)* is the number of residues of domain *a* that are located inside the radius of domain *b* and *vice versa*. The radius of a domain is the distance from the centroid that includes 90% of residues of the domain, in other words, the 90 percentile of the domain distance distribution.

Q is then given by

$$Q = w_1 Q_1 + w_2 Q_2, \qquad (15)$$

where w_1 and w_2 are weights (empirically we set $w_1 = 0.5$ and $w_2 = 0.5$).

For a given query protein, we carry out an exploration of the *N*-matrix at different threshold values starting from a low value. The exploration terminates when either all domains generated by SVD have less than 40 residues or the sum of residues involved in domains greater than 40 residues is less than two-third the number of residues in the query protein. The partition with the smallest *Q* score is selected as the solution.

Fig. 1

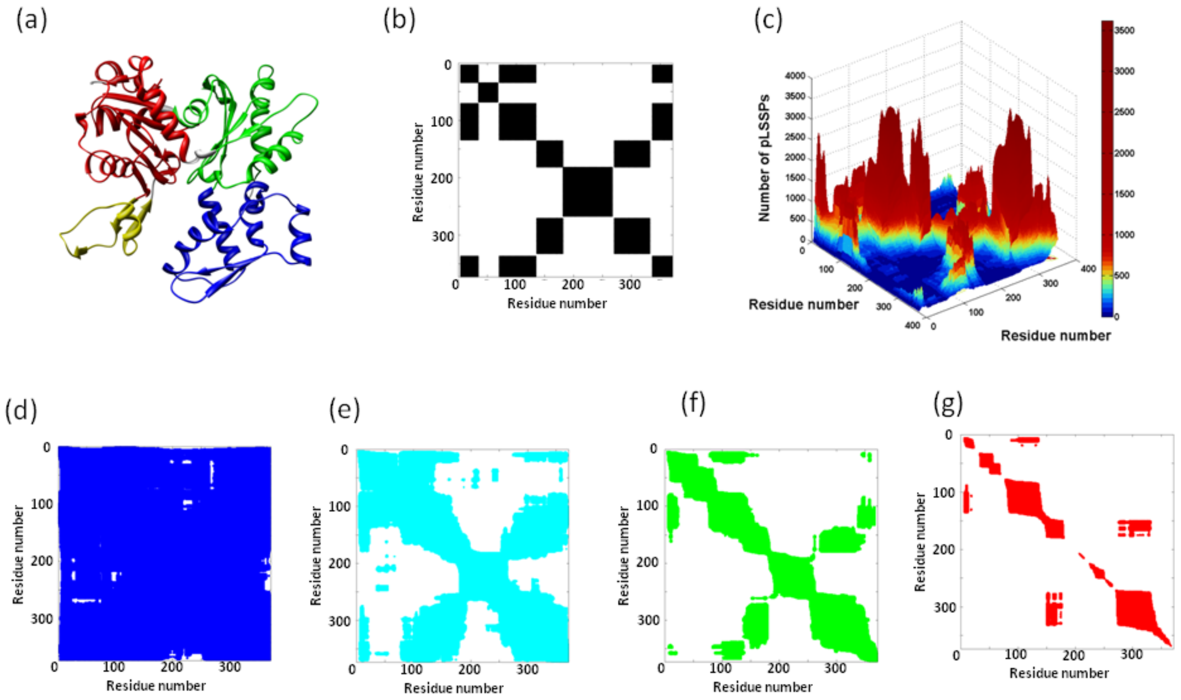

Fig. 1 legend: (a) Structure of 1atnA colored according to the CATH domains. (b) The *M*-matrix (matrix representation of an equivalence relation) using the CATH domain definition. (c) A 3D plot of the *N*-matrix. The values of the matrix elements are plotted along the z-axis, using colors according to the color scheme shown to the right of the graph. (d) to (g) Binarized *N*-matrices “sliced” at thresholds of 10, 100, 300 and 1000, respectively. The color of the matrix approximately corresponds to the color of the 3D plot of panel (c) at the slicing level.

Fig. 2

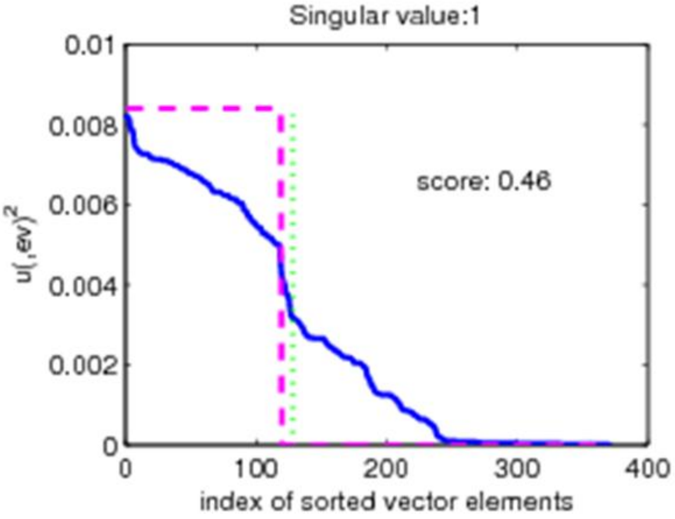

Fig.2 legend: The blue curve is the plot of the squares of the elements (sorted by decreasing order of magnitude) of the singular vector associated with the largest of the 4 singular values from the singular value decomposition of the binary matrix shown in Fig. 1f. The X-axis is for the elements of the singular vector, which correspond to the (scrambled) residues of 1atnA. The vertical dashed pink line is at the position of the singular value. In the ideal case, the elements to the left of the vertical line have the same non-zero value and correspond to the residues of the domain; all others are zero. The green dotted vertical line next to the pink one corresponds to a slight optimization of the singular value to find a position in the vicinity where the blue curve has the maximal slope. The score is the difference between the areas under the blue and the pink curves.