

# A new method for mapping macromolecular topography

Mihaly Mezei

*Department of Physiology and Biophysics, Mount Sinai School of Medicine, NYU, New York, NY 10029, USA.*

---

## Abstract

A new method, using circular variance, is introduced for mapping macromolecular topography. Circular variance, generally used to measure angular spread, can be used to characterize molecular structures based on a simple idea. It will be shown that the circular variance of vectors drawn from some origin to a set of points is well correlated with the degree to which that origin is inside/outside the chosen points. In addition, it has continuous derivatives that are also easy to compute. This concept will be shown to be useful for (i) distinguishing between atoms near the surface of a macromolecule and those in either the deep interior or remote exterior; (ii) identifying invaginations (even shallow ones) and (iii) detecting linker regions that interconnect two domains.

*Keywords:* circular variance; inside/outside; surface pockets; domain separation;

---

## 1. Introduction

Several geometrical properties of macromolecules are easy to characterize once visualized but hard to identify from atomic coordinates alone. Examples include whether a given atom is at or near the molecular surface, overall molecular shape, or determining whether a point (*e.g.*, a water molecule) is inside a cavity, outside the molecule or in a pocket (invagination).

The aim of this paper is to show that the concept of circular variance [1] — generally used in directional statistics (*i.e.*, characterizing angular spreads) — is invaluable in answering several such questions. The relation of angular spread to macromolecular topography can be visualized if we realize that a compact object can be illuminated either with a spotlight from a *distance* or with a light bulb placed *inside*.

Specifically, it will be shown that a circular variance can be used to establish a smooth scale for the relation between a query point and a set of points: one end of the scale represents the case of the query point being in the central region of the set while the other end of the scale corresponds to the case of the query point being way outside. The mid range of the scale corresponds to placements in the surface region.

Sec. 2.1. presents the formulae proposed and the intuitive argument leading to them; Sec. 2.2. gives an analytical demonstration of their validity for a model system; Sec. 2.3.

describes the test used for realistic systems; Sec. 2.4 discusses the relation of the method proposed to other methods and to methods for surface detection; Sec. 2.5. shows how the new method can be applied to the detection of surface pockets (even shallow ones); and Sec. 2.6. shows an other application, that of detecting linker regions. The software involved in these calculations is described in Sec. 2.7.; the results are presented and discussed in Sec. 3 and conclusions are drawn in Sec. 4.

## 2. Methods

### 2.1. Circular variance and its relation to the degree of ‘insideness’

The spread in a series of measurements is usually characterized by the standard deviation around the mean. Angular measurements, however, present a problem since  $0^\circ$  and  $360^\circ$  are equivalent. Circular variance was introduced to overcome this problem: for a set of angles  $\{\theta_i, i = 1, \dots, n\}$ , it is constructed to represent the angular range these angles span:

$$CV = 1 - \left[ \left( \sum_{i=1}^n \cos \theta_i \right)^2 + \left( \sum_{i=1}^n \sin \theta_i \right)^2 \right]^{1/2} / n. \quad (1)$$

When all the  $\theta_i$ 's are identical,  $CV=0$  (no spread) while for  $\theta_i$ 's uniformly distributed in the  $0^\circ$ – $360^\circ$  interval  $CV=1$  (maximum spread).  $CV$  has been used, for example, to characterize the spread of protein torsion angles [2].

An alternative, more general formulation for  $CV$  can be obtained as follows. Considering the representation of angles in the unit circle, the  $x$  and  $y$  components of the unit vector radius  $\mathbf{e}_i$  corresponding to the angle  $\theta_i$  are just  $\cos \theta_i$ , and  $\sin \theta_i$ , respectively. Thus the quantity inside the square brackets of Eq. (1) is just the square of the vectorial sum of the  $\mathbf{e}_i$ 's. This allows us to write  $CV$  as follows:

$$CV = 1 - \left| \sum_{i=1}^n \mathbf{e}_i \right| / n = 1 - \left| \sum_{i=1}^n \mathbf{e}_i \right| / \sum_{i=1}^n |\mathbf{e}_i|. \quad (2)$$

— the second equation follows since  $|\mathbf{e}_i| = 1$ . Eq. (2) is more general than Eq. (1) since it can be applied to unit vectors that don't lie in the same plane.

The fundamental idea of this paper is the use of the circular variance to characterize the relation between a point  $\mathbf{R}_o$  and a set of points  $\{\mathbf{r}_i\}$ . Applying Eq. (2) to the directions of vectors drawn from  $\mathbf{R}_o$  to  $\{\mathbf{r}_i\}$  gives the circular variance  $CV$  of these vectors as

$$CV = 1 - \left| \sum_{\{\mathbf{r}_i\}} [(\mathbf{R}_o - \mathbf{r}_i) / |(\mathbf{R}_o - \mathbf{r}_i)|] \right| / n \quad (3)$$

where  $n$  is the number of points in  $\{\mathbf{r}_i\}$ . When  $\mathbf{R}_o$  is infinitely far from  $\{\mathbf{r}_i\}$ , all vectors drawn are parallel, thus the vectorial and scalar sums coincide, resulting in  $CV=0$ ; when  $\mathbf{R}_o$  is in the ‘middle’ of  $\{\mathbf{r}_i\}$ , the vectorial sums tend to zero, resulting in  $CV=1$  (see Figure 1a). These arguments indicate that  $CV$  can provide a reasonable (albeit approximate) measure of how deep inside or how far outside  $\{\mathbf{r}_i\}$  the point  $\mathbf{R}_o$  is.

We also introduced a quantity  $CV^w$ , that will be called weighted circular variance, a generalization of the second equality in Eq. (1) whose properties are similarly to  $CV$ :

$$CV^w = 1 - \left| \frac{\sum_{\{\mathbf{r}_i\}} (\mathbf{R}_o - \mathbf{r}_i)}{\sum_{\{\mathbf{r}_i\}} |\mathbf{R}_o - \mathbf{r}_i|} \right|. \quad (4)$$

Unlike  $CV$ ,  $CV^w$  gives more weight to points farther from  $\mathbf{R}_o$  than to those in its immediate vicinity and its computation requires one less division per point. For one application it was found to be a better indicator than the original  $CV$ .

The choice of atoms included in the set  $\{\mathbf{r}_i\}$  provides an option for further refinements. The simplest tool for such selection is the introduction of a limit to the distances between the query point  $\mathbf{R}_o$  and the points chosen to be included in  $\{\mathbf{r}_i\}$ . This not only offers a significant reduction in the computing effort involved but also allows for the control of ‘smoothing’ the surface. Figure 1b shows an example of the effect of different cutoff radii: using the shorter cutoff (circle with broken lines) the points included will fall to one side of  $\mathbf{R}_o$  thus it will appear as being outside while the larger cutoff (circle with full line) will include points on both side of  $\mathbf{R}_o$ , indicating an interior position. The effect of this cutoff radius is therefore analogous to the effect of the solvent radius for the calculation of the solvent-accessible surface. An additional benefit of using a spherical cutoff is that it can also help to eliminate the skewing influence of the usually irregular macromolecular shape.

The computational effort involved in the calculation (complexity) of  $CV$  or  $CV^w$  for all atoms in a macromolecule of  $n$  atoms is only of  $O(n)$  when a fixed cutoff is used. Without using a cutoff, however, the computational effort is of  $O(n^2)$ .

## 2.2. Exact expressions for $CV$ and $CV^w$ in a model system

The qualitative arguments given above connecting  $CV$  and  $CV^w$  can be quantified using a model system. If we assume that the set of points under consideration is uniformly distributed within a sphere of radius  $r$  and  $CV^w$  is calculated within a sphere of radius  $R$  then  $CV^w$  can be analytically expressed as a function of the distance  $x$  between  $\mathbf{R}_o$  and the center of the set,  $\mathbf{O}$  (see Figure 2). For  $x \leq r$ :

$$CV^w(x, r, R) = 1 - \frac{\int_{r-x}^R \int_0^{\phi(r, \rho, x)} 2\pi \rho^3 \sin(\phi) \cos(\phi) d\rho d\phi}{\int_0^{r-x} 4\pi \rho^3 d\rho + \int_{r-x}^R \int_0^{\phi(r, \rho, x)} 2\pi \rho^3 \sin(\phi) d\rho d\phi} \quad (5)$$

where

$$\phi(r, \rho, x) = \cos^{-1} \frac{x^2 + \rho^2 - r^2}{2x\rho} \quad (6)$$

and for  $x > r$ :

$$CV^w(x, r, R) = 1 - CV_{\text{num}}^w(x, r, R)/CV_{\text{denom}}^w(x, r, R) \quad (7)$$

with

$$CV_{\text{num}}^w(x, r, R) = \int_{x-y(R,r,x)}^R \int_0^{\phi(R,\rho,x)} 2\pi\rho^3 \sin(\phi) \cos(\phi) d\rho d\phi + \int_{y(R,r,x)}^r \int_0^{\phi(r,\rho,x)} 2\pi(x - \rho \cos(\phi))\rho^2 \sin(\phi) d\rho d\phi \quad (8)$$

and

$$CV_{\text{denom}}^w(x, r, R) = \int_{x-y(R,r,x)}^R \int_0^{\phi(R,\rho,x)} 2\pi\rho^3 \sin(\phi) d\rho d\phi + \int_{y(R,r,x)}^r \int_0^{\phi(r,\rho,x)} 2\pi\rho^2 [x^2 + \rho^2 - 2x\rho \cos(\phi)]^{1/2} \sin(\phi) d\rho d\phi \quad (9)$$

where

$$y(R, r, x) = (x^2 + r^2 - R^2)/2x. \quad (10)$$

Eqs. (5) and (7) can be evaluated analytically resulting in rational fractions (*i.e.*, ratios of polynomials in  $x$ ,  $r$  and  $R$ ) [3]. The actual expressions are given in the Appendix. The numerators of Eqs. (5) and (7) are obtained after recognizing that pairing vectors at angles  $\phi$  with vectors at angles  $-\phi$  results in vectorial sums that always lie on the line connecting  $\mathbf{O}$  and  $\mathbf{R}_0$ .

Removing the factors representing the vector lengths in Eqs. (5-9) results in similar expressions for  $CV$ . This means that in the integrals (five out of seven) where  $\rho^3$  appears it has to be replaced by  $\rho^2$ , the factor  $[x^2 + \rho^2 - 2x\rho \cos(\phi)]^{1/2}$  has to be removed from Eq. (9) and the factor  $(x - \rho \cos(\phi))$  in Eq. (8) has to be replaced by

$$\cos(\phi_R) = (x - \rho \cos(\phi))/[x^2 + \rho^2 - 2x\rho \cos(\phi)]^{1/2}. \quad (11)$$

It is also instructive to examine the special case of  $x = r$  and  $r = R/c$  in the limit of  $c \rightarrow 0$ . As  $c$  approaches zero, the boundary approaches an infinite plane, thus it is expected that  $CV$  and  $CV^w$  will tend to 1/2 when  $\mathbf{R}_0$  is at the boundary. In this special case the expression for Eq. (6) and its counterpart for  $CV$  simplifies to

$$CV = (20 - 3c^2)/(40 + 15c) \quad \text{and} \quad CV^w = (c^2 - 30)/(24c - 60) \quad (12)$$

verifying this intuitive notion.

Figure 3 shows  $CV$  and  $CV^w$  as a function of  $x$  for different cutoff radii. The radius of the sphere containing the set of points,  $r$ , was kept at 20 — it is the ratio  $R/r$  that gives rise to different shaped curves. Both functions monotonically decrease as  $x$  increases (*i.e.*, as the test point moves from the interior to the exterior) demonstrating that the intuitive notion connecting  $CV$  and  $CV^w$  with the degree of ‘insideness’ is verified without exception

on this model system. It is also seen that  $CV^w$  is closer to being linear than  $CV$  (especially in the ‘inside’ part of the curves), but  $CV$  varies more steeply in the boundary region.

## 2.2. Studies of realistic systems

In the present work, the theoretical arguments given above (Eqs. (5-12)) will be supplemented with calculation on actual molecules.  $CV$  and  $CV^w$  will be calculated for several proteins employing different cutoffs and the result will be compared with a geometrically derived residue depth calculated as follows. The fraction of the atomic surface exposed to the outside of a molecule can be considered to be a measure of the position of that atom in the boundary: the atom with the highest fraction of exposed surface can be considered the outermost. Thus, ‘shaving’ the macromolecule atom by atom via removal of the most exposed atom in each step results in an ordering of the atoms by their closeness to the molecular surface.

Furthermore, based on the order the atoms were shaved off, they were assigned to layers of ca 1 Å width as follows. Assuming spherical shape and 1 Å layer thickness the number of such layers can be deduced from the volume of the protein. Assuming uniform distribution, the number of atoms  $n_i$  in each layer  $i$  can be obtained by appropriate partitioning of the set of atoms. The first  $n_1$  atoms shaved were assigned to layer 1, the next  $n_2$  layers to layer 2, etc. The actual thickness of these layers will be less than one since the proteins under consideration are not perfectly spherical.

## 2.3. Relation of the circular variance measure to surface detection algorithms

Considerable attention has been paid to the problem of finding the surface atoms of a macromolecule. Recent work by Deanda and Pearlman compared a number of approaches to surface detection and atomic surface area determination based on different definitions of surface and different indicators [4]. Analytical approaches, based on geometric concepts (*e.g.*, convex hull, Voronoi and/or Delaunay tessellation) have also been invoked as reviewed by Edelsbrunner *et al* [5].

Our measure of the relation of a point to the atoms of a macromolecule suggests that there is a range of  $CV$  or  $CV^w$  values that is diagnostic of the surface region — in the limit of our model system it was found to be 0.5 for both. One of the techniques discussed by Deanda and Pearlman, called SOV (for Sum Of Vectors) attempts to correlate the surface area of an atom with a quantity that is the numerator of our Eq. (4) defining  $CV^w$  (using a short cutoff). They found that SOV fails to predict reliably the contribution of an atom to the surface area. While our definition of  $CV$  and  $CV^w$  and the use of larger cutoffs likely factors out some of the influence of the specifics of the local environment, the correlation between our  $CV$ ’s and the surface area of an atom is still only approximate (as expected).

However, the information content of  $CV$  or  $CV^w$  is primarily related to the depth of the test point. They not only answer the question if an atom is in the surface *region* or not, with the same effort they also give a measure of its likely distance from the surface. Thus, it is more relevant to compare it with measures that indicate the depth of the point (see, *e.g.*, Chakravarty and Varadarajan [6]). Also, by varying the cutoff radius for the selection of

atoms to be included in the set  $\{\mathbf{r}_i\}$  surface features can be smoothed out — an additional indication that the link between  $CV$  and  $CV^w$  is, partly by design, only approximate.

#### 2.4. Identification of surface pockets

Identification of surface pockets (invaginations) of macromolecules is also an area of active research. Besides being of intrinsic interest, such algorithms are relevant to searches for binding sites and their analyses. Recently, Edelsbrunner *et al.* [5] have shown that the Delaunay tessellation defined by the atoms of a macromolecule can be conveniently used not only to detect cavities, but also to detect surface pockets, as long as the opening of the pocket is narrower than some cross section of the pocket’s inside. However, not all binding sites have this property, thus the problem of identifying these shallower pockets remains. Mitchell, Kerr, and Ten Eyck [7] introduced two methods for rapid characterization of molecular shape based on the density of atoms near the molecular surface. Their shape descriptors can also be used to find pockets on the molecular surface.

Stahl *et al.* [8] introduced a grid-based technique for the detection of surface pockets. Since, as discussed above, circular variance can distinguish between points inside and outside a pocket it can be incorporated into the procedure of Stahl *et al.* as follows:

1. Overlay a grid (whose spacing is less than the size of an atom) on the macromolecule and remove the gridpoints that are covered by an atom of the macromolecule.
2. Find the connected clusters of the remaining gridpoints (assuming that only neighboring gridpoints are connected). One of these clusters (the largest) will include all the gridpoints external to the macromolecule while the rest of the clusters (if any) will delineate various cavities.
3. Calculate the  $CV$  for the gridpoints in the external cluster with respect to the atoms of the macromolecule, using a cutoff radius that is larger than the maximum width of a pocket to be considered.
4. Eliminate all gridpoints where the value of  $CV$  is below a threshold value,  $CV^{\max}$ .
5. Again, separate the remaining gridpoints into connected clusters — each of these clusters will represent one pocket. The more gridpoints are in the cluster, the larger the pocket is.

The critical part of the algorithm is at steps 3 and 4, the decision deciding if a gridpoint is in a pocket or not. This is exactly the point where procedure proposed here differs from that of Stahl *et al.*. Using the circular variance to decide if a gridpoint is in a cavity has the advantage that it lets one decide if a gridpoint is deep in a cavity or just at the surface *without the consideration of other gridpoints*. Further, the choice of cutoff radius allows a simple control over the maximum width of the cavities detected. If the cutoff radius is less than the half of the diameter of the invagination’s opening, it will not be considered a pocket.

Assuming that the number of gridpoints is proportional to the number of atoms,  $n$ , the computational effort involved in each step of the proposed procedure is of  $O(n)$  (with the use of a cutoff), thus the search for pockets can be done in  $O(n)$  steps. The assumption

about the number of gridpoints is reasonable since it is proportional to the volume of the cube enclosing the macromolecule, which is proportional to the number of atoms. A possible speedup from  $O(n)$  to  $O(n^{2/3})$  would result if instead of a three-dimensional grid a few layers of dot-surfaces were used in a similar fashion (R. Pearlman, personal communication). As a tradeoff, the number of points in each cluster representing a pocket would provide an estimate for the surface of the pocket instead of its volume.

There are two technical issues related to the grid used in the pocket detection: the position and orientation of the macromolecule in the grid. First, due to the finite spacing of the grid the set of gridpoints not covered by the macromolecule may vary in size and/or shape. This uncertainty can be reduced by using smaller grid spacing and/or repeating the calculation with the macromolecule slightly shifted. Second, given a fixed number of gridpoints, the orientation of the macromolecule affects the spacing of the grid: the orientation with the smallest enclosing rectangle will use the smallest grid spacing possible for a given number of gridpoints [9].

### 2.5. Detection of domain separation

Recent work by Anselmi *et al.* [10] presented a theoretical method for the identification of protein domains through the topological analysis of certain geometrical properties of the protein backbone chain. However, the fact that a linker region connecting two domains falls outside both domains suggests that the concept of circular variance can also be exploited for this purpose. To that effect, the concept of circular variance map is introduced for a macromolecule of  $N$  residues as an  $N \times N$  matrix,  $\mathbf{CV}$ , whose elements are defined as

$$CV_{i,j} = 1 - \left[ \sum_{k \in (j,i]} \mathbf{r}_{i,k} / |\mathbf{r}_{i,k}| \right] / |i - j| \quad (13)$$

together with its weighted variant,  $\mathbf{CV}^w$ :

$$CV_{i,j}^w = 1 - \left| \sum_{k \in (j,i]} \mathbf{r}_{i,k} \right| / \sum_{k \in (j,i]} |\mathbf{r}_{i,k}| \quad (14)$$

where  $\mathbf{r}_{i,k}$  is the vector pointing from residue  $k$  to residue  $i$ . The position of a residue is determined by one of its ‘representative’ atoms — for proteins the natural choice is the  $\alpha$  carbon. The calculation and plotting of circular variance maps defined by The computational effort involved in generating a  $CV$  map is not worse than of  $O(N^2)$  since the sums involved in  $CV_{i,j}$  are part of the sums forming  $CV_{i,j+1}$ , allowing for a recursive construction of the maps. Since a linker region is outside the domains it separates, the map entries corresponding to the linker residue columns will all have  $CV$  values close to zero. Thus a vertical swath of low  $CV$  values in the map indicates a domain-separating linker region.

## 2.6. Software implementations

The calculation of circular variances for the atoms of a macromolecule as well as for the solvent molecules around it with respect to the atoms of the macromolecule has been implemented in the program Simulaid [11]. The program GEPOL93 [12] was modified to execute the shaving process and perform the comparison with the circular variances.

The orientational optimization within the grid (analogous to the orientational optimization aimed at maximizing the distances between periodic images for a simulation using periodic boundary conditions [13]) was also performed with the program Simulaid [11]. It should be stressed that finding the orientation of a macromolecule with the smallest enclosing rectangle is relevant to any numerical procedure based on a grid (*e.g.*, Poisson-Boltzmann solvers). The generation of the grids required for the detection of pockets took advantage of the cavity-biased grand-canonical ensemble code of the Monte Carlo program MMC [14] and thus the CV-based pocket detection has been implemented in that program. Eqs. (13,14) for the detection of linker regions have also been implemented in the program Simulaid [11].

Both Simulaid and MMC are written in Fortran-77 and have been successfully run on several platforms (including Linux). Simulaid has an optional graphics interface using Iris-GL. Both programs are available at the URL <http://inka.mssm.edu/~mezei>.

## 3. Results and discussion

The ability of  $CV$  and  $CV^w$  to characterize the degree of burial of the atoms was studied on six proteins (PDB ids: 1amm [15], 2fok [16], 5tim [17], 1cwq [18], 1bpi [19], 2de8 [20]). Despite the smallness of the sample it was found to be adequate since the results were uniform for all the proteins studied.

For the test, each protein underwent the ‘shaving’ process described in Sec. 2.3. Table 1 shows four different correlations between of the calculated  $CV$  values and the layer number the atom belongs to, using the longest cutoff employed in this study, 10 Å: 1) correlation between all layer numbers and individual  $CV$  values; 2) correlation between all layer numbers that correspond to depths less than the cutoff and individual  $CV$  values; 3) correlation between all layer numbers and  $CV$  values averaged in that layer; 4) correlation between all layer numbers that correspond to depths less than the cutoff and  $CV$  values averaged in that layer. Since  $CV$  can not discriminate beyond the cutoff distance used, the correlations including all points are somewhat weaker. The correlations using the layer averaged  $CV$  values are much stronger — for the longer cutoff values they reach one. This reflects the fact that there is a certain amount of arbitrariness in the determination of the layer number. The last column of the Table 1 gives the average difference between  $CV$  and  $CV^w$  — they are small, indicating that  $CV$  and  $CV^w$  both are similarly good indicators of the degree of ‘insiderness’.

Table 2 shows the cutoff dependence of the correlations between the layer number and the  $CV$  values. The steady improvement with the cutoff was also seen with all the other correlations shown in Table 1. The magnitude of the improvement was also similar, except

for the correlations between the limited layer numbers and  $\langle CV \rangle$ , where it was much less, about 10%. The increase of correlation found even in the restricted data set indicates that the expected cancellations in the vectorial sum occur to a much better degree with the larger cutoff, suggesting the use of a minimum of 6 Å.

Figure 4 shows the means of  $CV$  for each layer for the largest of the proteins studied (2fok) calculated with 10 Å cutoff. The values saturate beyond layer 13 at  $CV \simeq 0.95$ . The fact that the saturation occurs ca 30% deeper than what would be expected if the layer thickness were the ideal 1 Å suggests that the actual layer thickness is about 0.7 Å. The deviation of the saturated value from 1.0 is an additional indication of the accuracy of  $CV$  as a measure of depth. The 0.05 deviation is indeed consistent with the calculated standard deviations.

The correlations between the atomic surface and  $CV$  were also calculated — they are shown in Table 3. As discussed earlier, these correlations are much weaker than the correlations between  $CV$  and the layer number. Furthermore, while  $CV$  correlates better with the layer number when using larger cutoffs, the correlations between  $CV$  and the surface area further weaken, showing that, as discussed earlier,  $CV$  is not a surface measure *by design*.

As a typical example, a 6 Å slice (perpendicular to the helix axes) of bacteriorhodopsin [18] (PDB id: 1cwq) is shown on Figure 5, extracted from a simulation of bacteriorhodopsin in a water bath for the study of internal hydration (S. Yancopoulos, M. Mezei, E. Mehler, and R. Osman, in preparation). Both the protein atoms and water molecules have been color-coded with their circular variance (calculated on the full structure with a 6 Å cutoff). Water spheres are reduced by 70% for clarity. The gradual decrease of  $CV$  is clearly discernible as one goes from the interior of the protein to the exterior region.

Having a single scalar quantity that represents the extent of burial of an atom in a macromolecule can find many application. For example, it can be incorporated into various models of interatomic interactions — a likely application of such application could be calculations of  $pK_A$ 's. Such developments are aided by the fact that the derivatives of  $CV$  and  $CV^w$  with respect to the coordinates of the test point  $\mathbf{R}_0$  as well as with respect to the coordinates of the points  $\mathbf{r}_i$  in the set can be computed with the same complexity as  $CV$  and  $CV^w$ . Furthermore, these derivatives are continuous when no cutoff is employed or can be made continuous by the replacement of the simple cutoff with a cutoff range where the contributions are smoothly scaled to zero (a computational device widely used in molecular dynamics simulations [21]).

The algorithm described in Sec. 2.5. for the determination of surface pockets has been tested successfully on several proteins. Figure 6 shows a typical result of the pocket finding calculations for the protein bromodomain [22] (PDB id: 1b91): the pockets calculated with a cutoff radius of 10 Å keeping all gridpoints with  $CV > 0.65$ . The grid spacing was 0.6 Å in the  $x$  and  $z$  directions, 0.7 Å in the  $y$  direction.

The success in finding the pockets on the protein surface opens the way to different important applications. Perhaps the most important one would be the search for binding sites — one of the bromodomain pocket found by this procedure is indeed its binding site.

Naturally, not all pockets are binding sites so being able to find pockets would be just one part of the search. Other applications may include search for interfaces where the existence of pockets and complementary ‘knobs’ are a likely indicators of surface complementarity — such ‘knobs’ could also be delineated by looking for clusters of low  $CV$  atoms (M. Filizola, personal communication). Again, such analyses would be part of the process only since surface complementarity is only a necessary condition for interface formation.

Figure 7 shows a typical example comparing the original and weighted circular variance maps using the protein  $\gamma$ B-crystallin [15] (PDB id: 1amm) — a dumbbell-shaped protein whose two domains are linked by residues 80-85 (also studied by Anselmi *et al.* [10]). Both maps clearly show the low- $CV$  swath at the site of the linker, but it is more prominent on the weighted map, indicating that  $CV^w$  is more sensitive than  $CV$  in this context. Similar differences were found on several other multi-domain proteins (not shown). An other example is shown on Figure 8, the weighted circular variance map of the 7-helix bacteriorhodopsin [18]. The black bars under the row averages span the residue ranges of the seven helices showing that the helix linker regions correspond to the low- $CV^w$  columns. Note also that while the row average of  $CV^w$  is low between helices 3 and 4, the maximum  $CV^w$  value is larger than in the other linker regions. This corresponds to the fact that helices 3 and 4 are shorter than the others thus this linker region is not quite outside the domains it separates.

In addition to the occasional vertical swaths indicating domain-separating regions, the  $CV$  maps have a fairly rich structure. This suggests an other potential use: characterization of the fold of the protein. If such characterization is found to be discriminating enough, it can be used for fold recognition. This avenue is planned to be explored in future work.

## 4. Conclusion

The intuitively deduced relation between the circular variance and depth of an atom in a macromolecule (or its distance from it) has been verified theoretically on a model system and demonstrated with numerical examples using a set of diverse proteins. Furthermore, it was shown that exploiting this relation circular variance can be a powerful tool for mapping macromolecular topography.

It is expected that more applications will follow the ones presented here. Of these future application of particular interest is the incorporation of circular variance into modeling of intermolecular interactions, as it possesses easily computable and continuous derivatives. Other possible avenues of investigation include detecting surface complementarities and the use of the circular variance maps for fold characterization.

## Acknowledgement

Critical comments on the manuscript by Professor George Rose are gratefully acknowledged. Juan Luis Pascual-Ahuir is thanked for permission to use the GEPOL93 [12] source code.

## References

- [1] K.V. Mardia, and P.E. Jupp., *Directional Statistics*, Wiley, New York, 1999.
- [2] M.W. MacArthur, and J.M. Thornton, *Proteins* 17 232 (1993).
- [3] S. Wolfram, *Mathematica, A System for Doing Mathematics by Computer*, Addison-Wesley, 1991.
- [4] F. Deanda, and R.S. Pearlman, *Macromolecules. J. Mol. Graphics and Modell.* 20 (2002) 415.
- [5] H. Edelsbrunner, M. Facello, and J. Liang, *On the Definition and Construction of Pockets in Macromolecules. Disc. Appl. Math.* 20 (1998) 83.
- [6] S. Chakravarty, and R. Varadarajan, *Structure Fold. Des.* 7 (1999) 723.
- [7] J.C. Mitchell, R. Kerr, and L.F. Ten Eyck, *J. Mol. Graphics and Modell.* 19 (2001) 325.
- [8] M. Stahl, C. Taroni, and G. Schneider, *Prot. Engnrng.* 13 (2000) 83.
- [9] M. Mezei, *Information Newsletter for Computer Simulation of Condensed Phases, CCP5, Daresbury Lab., No 47, (2000).*
- [10] C. Anselmi, G. Bocchinfuso, A. Scipioni, and P. De Santis, *Proteins* 58 (2001) 218.
- [11] M. Mezei, *Simulaid, Simulation setup utilities*; URL: <http://inka.mssm.edu/~mezei/simulaid>
- [12] J.L. Pascual-Ahuir, I. Tuñon, and E. Silla, *J. Comput. Chem.* 15 (1994) 1127.
- [13] M. Mezei, *J. Comput. Chem.* 18 (1997) 812.
- [14] M. Mezei, *MMC, Monte Carlo program for the simulation of molecular assemblies*; URL: <http://inka.mssm.edu/~mezei/mmc>
- [15] V.S. Kumaraswamy, P.F. Lindley, C. Slingsby, and I.D. Glover, *Acta Crystallogr D Biol Crystallogr.* 52 (1996) 611.
- [16] D.A. Wah, J. Bitinaite, I. Schildkraut, and A.K. Aggarwal, *Proc. Nat. Acad. Sci. USA* 95 (1998) 10564.
- [17] R.K. Wierenga, M.E.M. Noble, G. Vriend, S. Nauche, and W.G.J. Hol, *J. Mol. Biol.* 220 (1991) 995.
- [18] H.J. Sass, J. Berendzen, D. Neff, R. Gessenich, P. Ormos, and G. Bueldt, *Nature* 406 (2000) 649.
- [19] S. Parkin, B. Rupp, and H. Hope, *Acta Crystallogr. D Biol. Crystallogr.* 52 (1996) 18.
- [20] C.D. Mol, T. Izumi, S. Mitra, and J.A. Tainer, *Nature* 403 (2000) 451.
- [21] F.H. Stillinger, and A. Rahman, *J. Chem. Phys.* 60 (1974) 1545.

[22] C. Dhalluin, J.E. Carlson, L. Zeng, C. He, A.K. Aggarwal, and M.M. Zhou, *Nature* 399 (1999) 491.

This article was published in the *Journal of Molecular Graphics and Modeling*, **Vol. 21**, M. Mezei, A new method for mapping macromolecular topography, pp 463-472, Copyright (2003), and is posted with permission from Elsevier.

## Appendix

The integrals in Eqs. (5) and (7) evaluate to the following expressions [3]:

For  $x \leq r$ :

$$CV^w(x, r, R) = 1 - \frac{-5 \left( r^2 - R^2 - 2rx + x^2 \right)^2 \left( r^2 - R^2 + 4rx + x^2 \right)}{4x \left( 4r^5 + 6R^5 - 15r^4x - 15R^4x + 20r^3x^2 + 10R^3x^2 + x^5 - 10r^2 \left( R^3 + x^3 \right) \right)} \quad (A1)$$

For  $x > r$ :

$$CV^w(x, r, R) = 1 - \frac{CV_{\text{num}}^w(x, r, R)}{CV_{\text{denom}}^w(x, r, R)} \quad (A2)$$

where

$$CV_{\text{num}}^w(x, r, R) = \left( \pi \left( r^2 - R^2 - 2rx + x^2 \right)^2 \left( r^2 - R^2 + 4rx + x^2 \right) \right) / (24x^2) \quad (A3)$$

and

$$CV_{\text{denom}}^w(x, r, R) = (\pi/480) \left( 240R^4 - (160R^3(-r^2 + R^2 + x^2))/x + 5(-r^2 + R^2 + x^2)^4/x^4 + \right. \\ \left. 16 \frac{-4rR^4R + 4R^4Rx(x-r)(4r^5 - 15r^4x + 20r^3x^2 - 10r^2x^3)}{x(-r+x)} + \right. \quad (A4) \\ \left. (-r^2 + R^2 + x^2) \times \right. \\ \left. \frac{1.1x^8 - (r^8 + R^8)/2 - 2R^6x^2 - 3R^4x^4 - 2R^2x^6 + 2r^6(R^2 + x^2) - 3r^4(R^2 + x^2)^2 + 2r^2(R^2 + x^2)^3}{x^5(-r^2 + R^2 + x^2)/(10x^2)} \right)$$

Table 1

Different correlations between layer number and  $CV$  calculated with  $R_c = 10 \text{ \AA}$ 

	1amm	2fok	5tim	1cwq	1bpi	1de8	$\langle CV^w - CV \rangle$
layer # - $CV$	0.88	0.83	0.84	0.82	0.88	0.84	-0.001
limited layer# - $CV$	0.89	0.86	0.87	0.82	0.88	0.86	-0.041
layer # - $\langle CV \rangle$	0.88	0.84	0.81	0.93	0.98	0.86	0.031
limited layer # - $\langle CV \rangle$	0.99	0.99	0.99	0.97	0.98	0.99	0.001

Table 2

Correlations between layer number and  $CV$  as a function of the cutoff

$R_c$	1amm	2fok	5tim	1cwq	1bpi	1de8
4	0.50	0.34	0.35	0.59	0.63	0.42
6	0.76	0.68	0.72	0.75	0.84	0.72
8	0.84	0.79	0.79	0.81	0.87	0.80
10	0.88	0.83	0.84	0.82	0.88	0.84
$\langle CV^w - CV \rangle$	0.00	0.00	0.00	-0.01	-0.01	0.00

Table 3

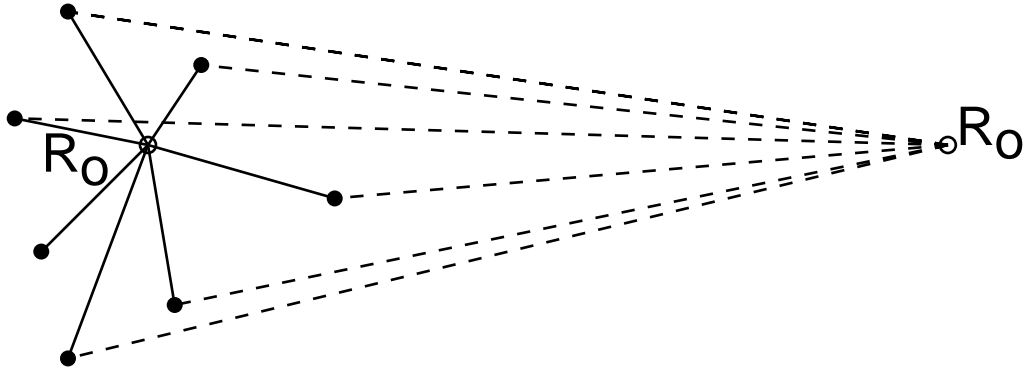
Correlations between exposure fraction and  $CV$  as a function of the cutoff

$R_c$	1amm	2fok	5tim	1cwq	1bpi	1de8
4	0.83	0.77	0.79	0.82	0.86	0.78
6	0.77	0.78	0.79	0.80	0.80	0.79
8	0.67	0.68	0.69	0.71	0.73	0.70
10	0.59	0.60	0.62	0.63	0.66	0.62
$\langle CV^w - CV \rangle$	-0.07	-0.06	-0.06	-0.07	-0.08	-0.06

## Figure captions

1. Illustration of the heuristics explaining the properties of circular variance. a) the two extreme situations; b) the effect of the cutoff. Filled circle represent the points of the set and open circles the query point  $\mathbf{R}_o$ .
2. The variables in the integrals giving  $CV^w$  for points uniformly distributed in a sphere and using a spherical cutoff; a:  $x \leq r$ ; b:  $x > r$ .
3.  $CV$  and  $CV^w$  as a function of  $x$  (the distance of the test point  $\mathbf{R}_o$  from the center of the sphere). Full line:  $CV^w$ ; broken line:  $CV$ . Plots for cutoff radii  $R = 20, 18, \dots, 2$  are shown; the radius of the sphere enclosing the set was set to 20.
4. Mean circular variance  $\langle CV \rangle$  as a function of the layer number, calculated with 10 Å cutoff for restriction endonuclease FokI (PDB id: 2fok). Error bars represent one standard deviation.
5. Variation of the circular variance as a function of ‘insiderness’ in a 6 Å slice of a snapshot taken from a simulation of bacteriorhodopsin (PDB id: 1cwq) in water bath. Atoms are color-coded by circular variance. Water molecules are shown with reduced size.
6. Pockets of the bromodomain (PDB id: 1b91). Gridpoints in pockets are shown as small green spheres, atoms have their usual color-code (H: white, C: gray; O: red; N: blue; S: yellow).
7. Comparison of the original and weighted circular variance maps for the protein  $\gamma$ B-crystallin (PDB id: 1amm). Color strip under each map shows the column averages.
8. Weighted circular variance map of bacteriorhodopsin. Color strip under the map shows the column averages. Black horizontal bars show the positions of the seven helices.

a



b

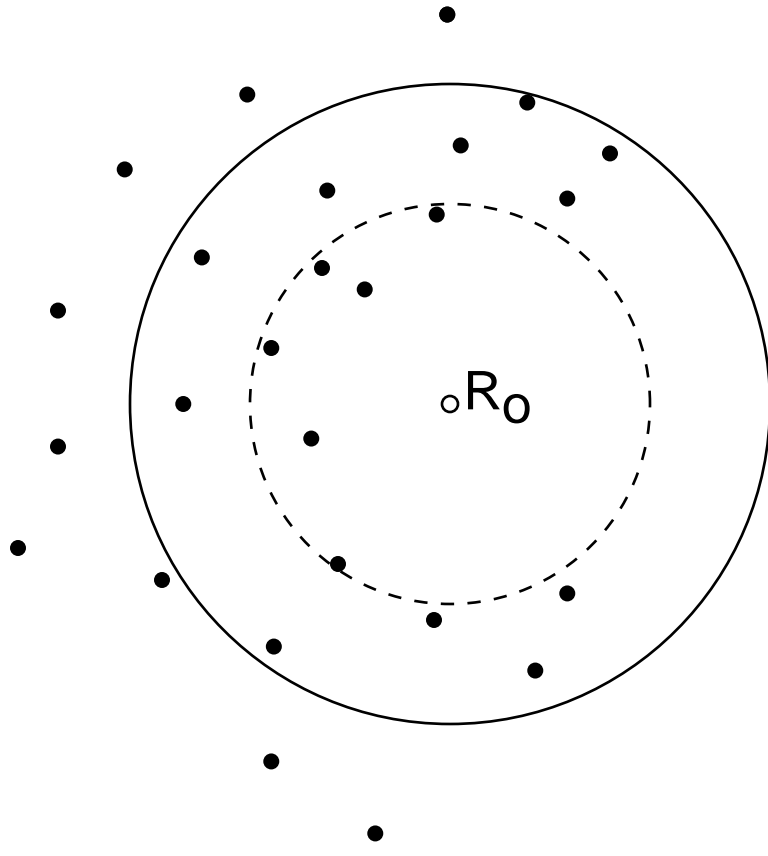
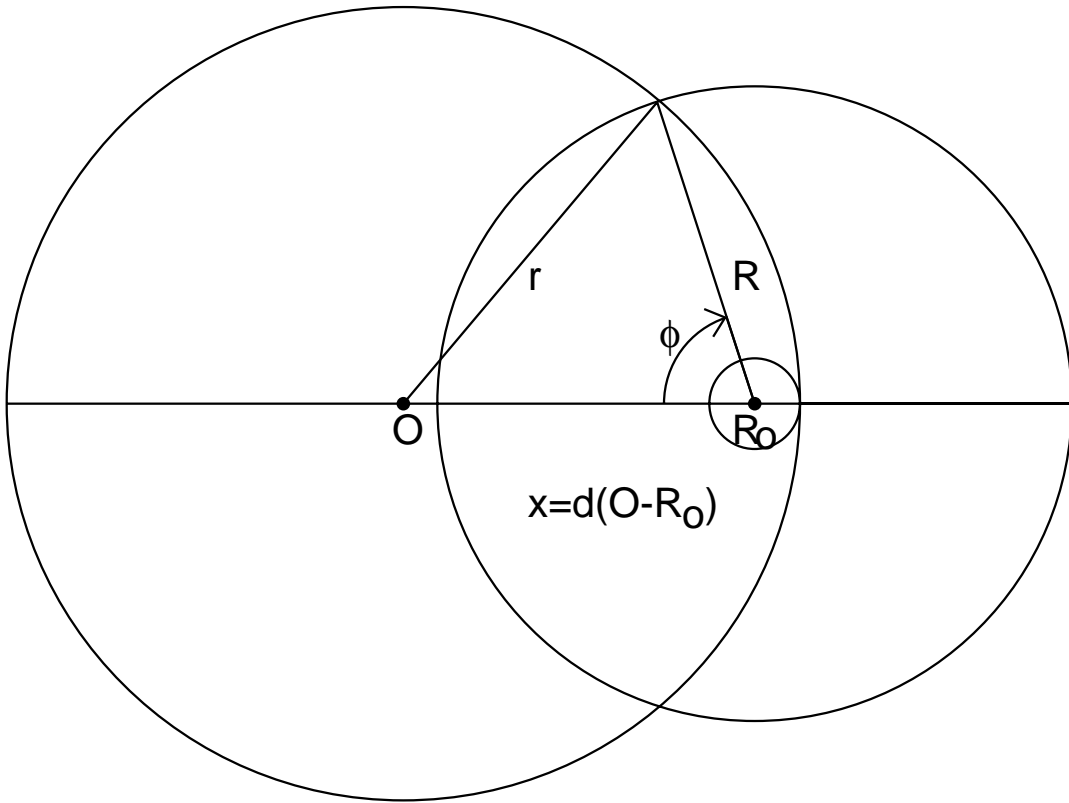


Figure 1

a



b

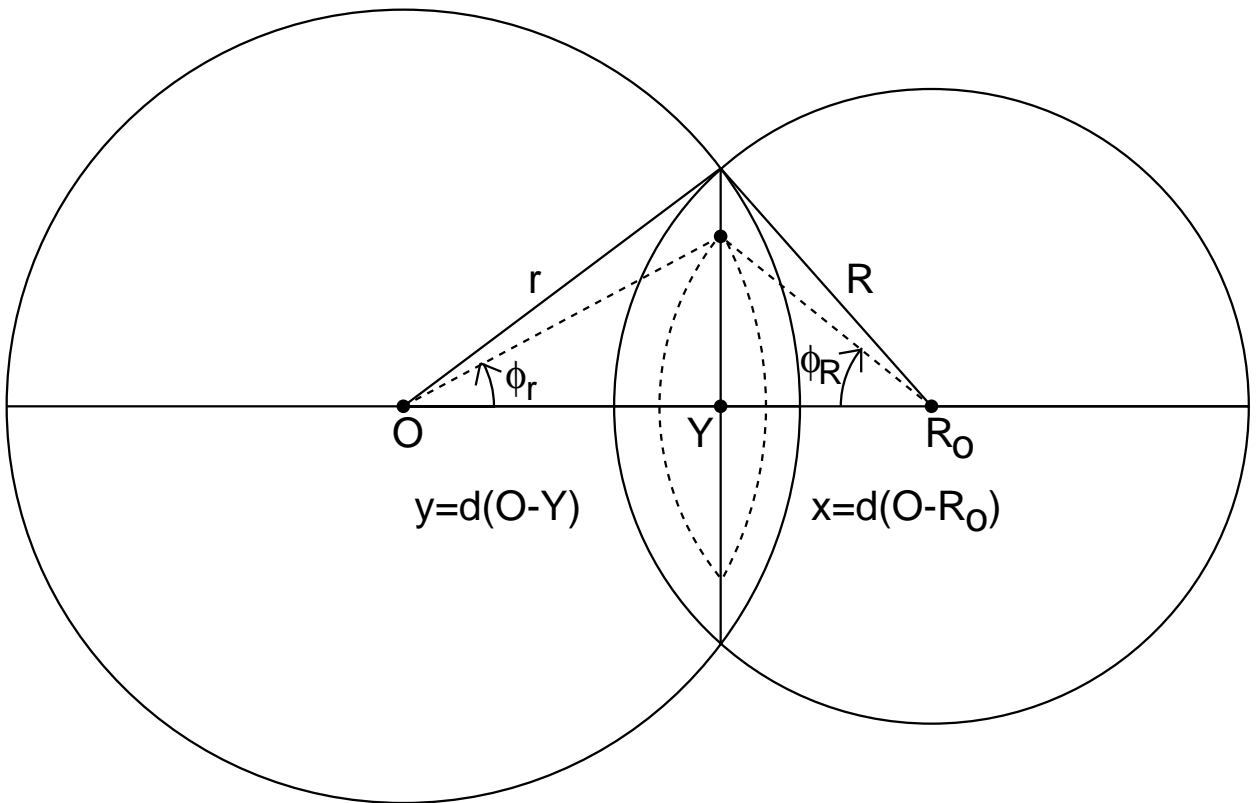
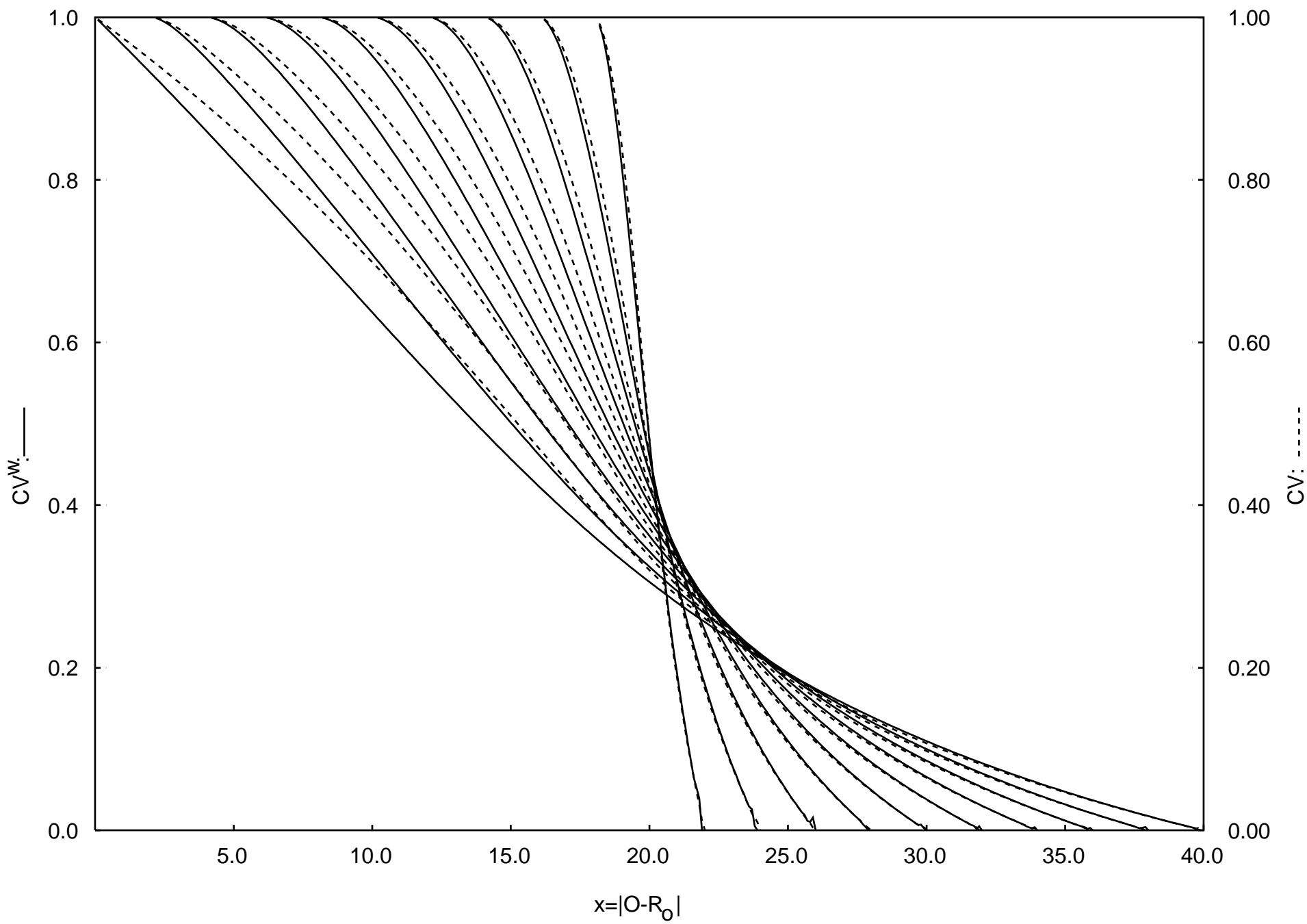
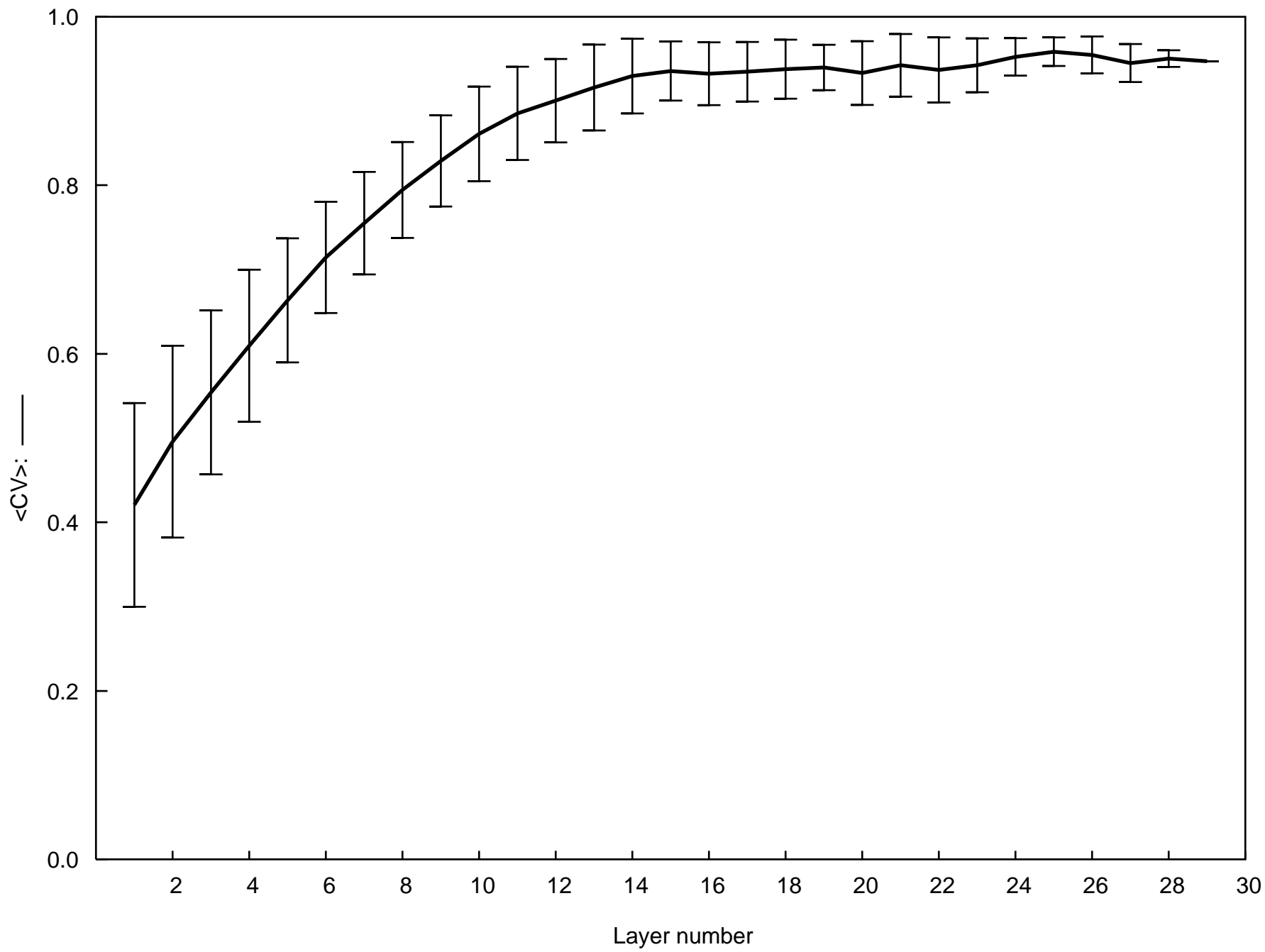
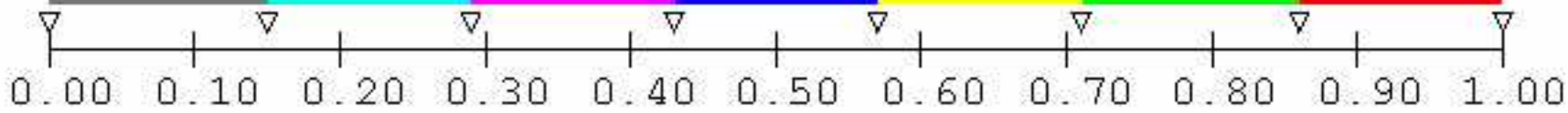
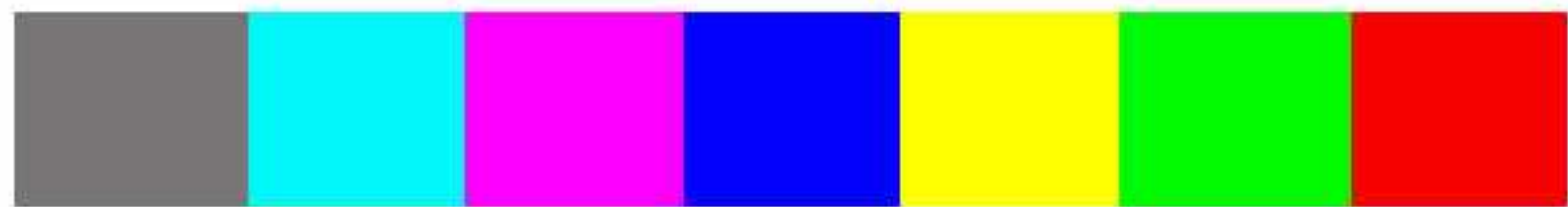
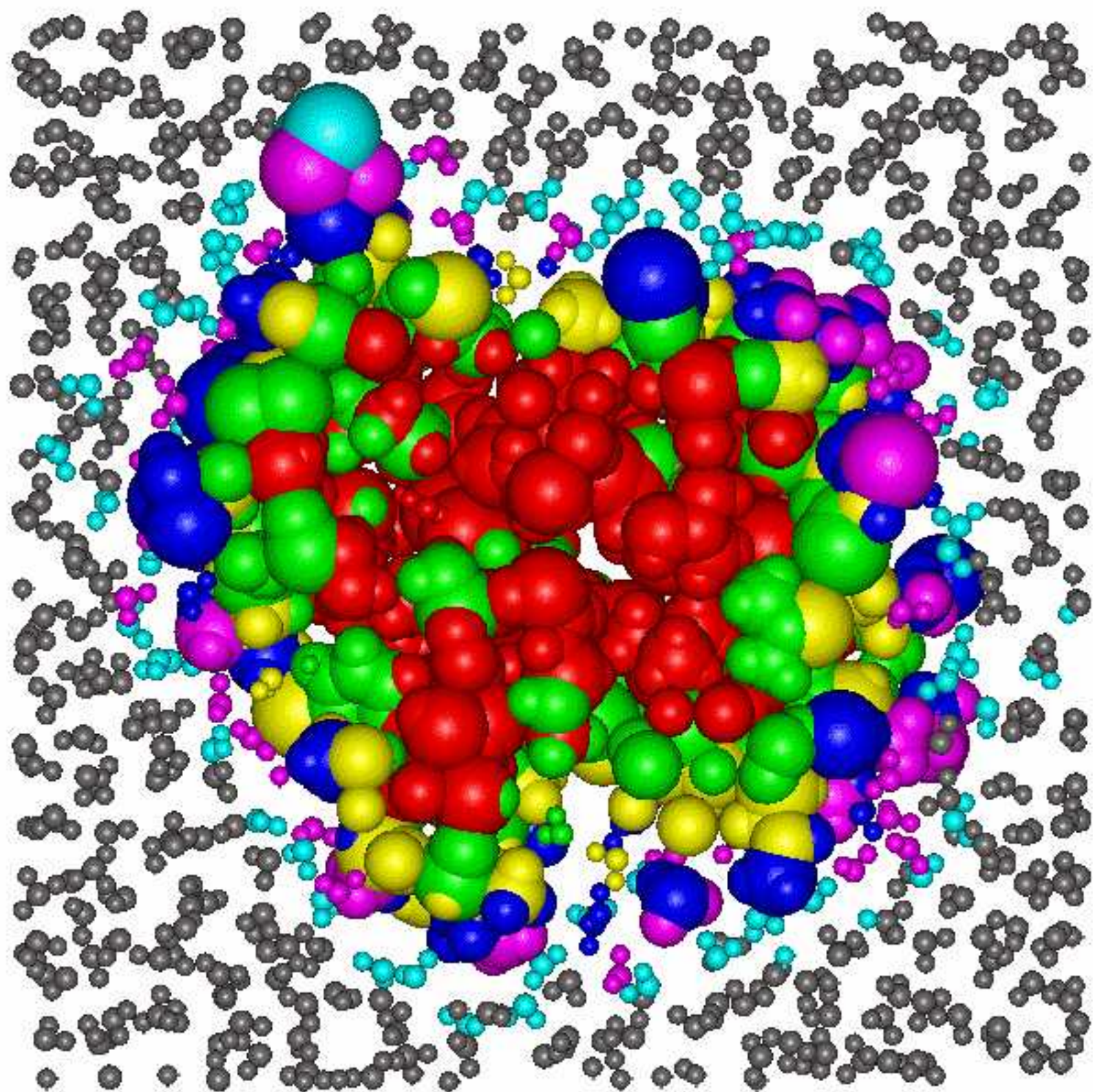
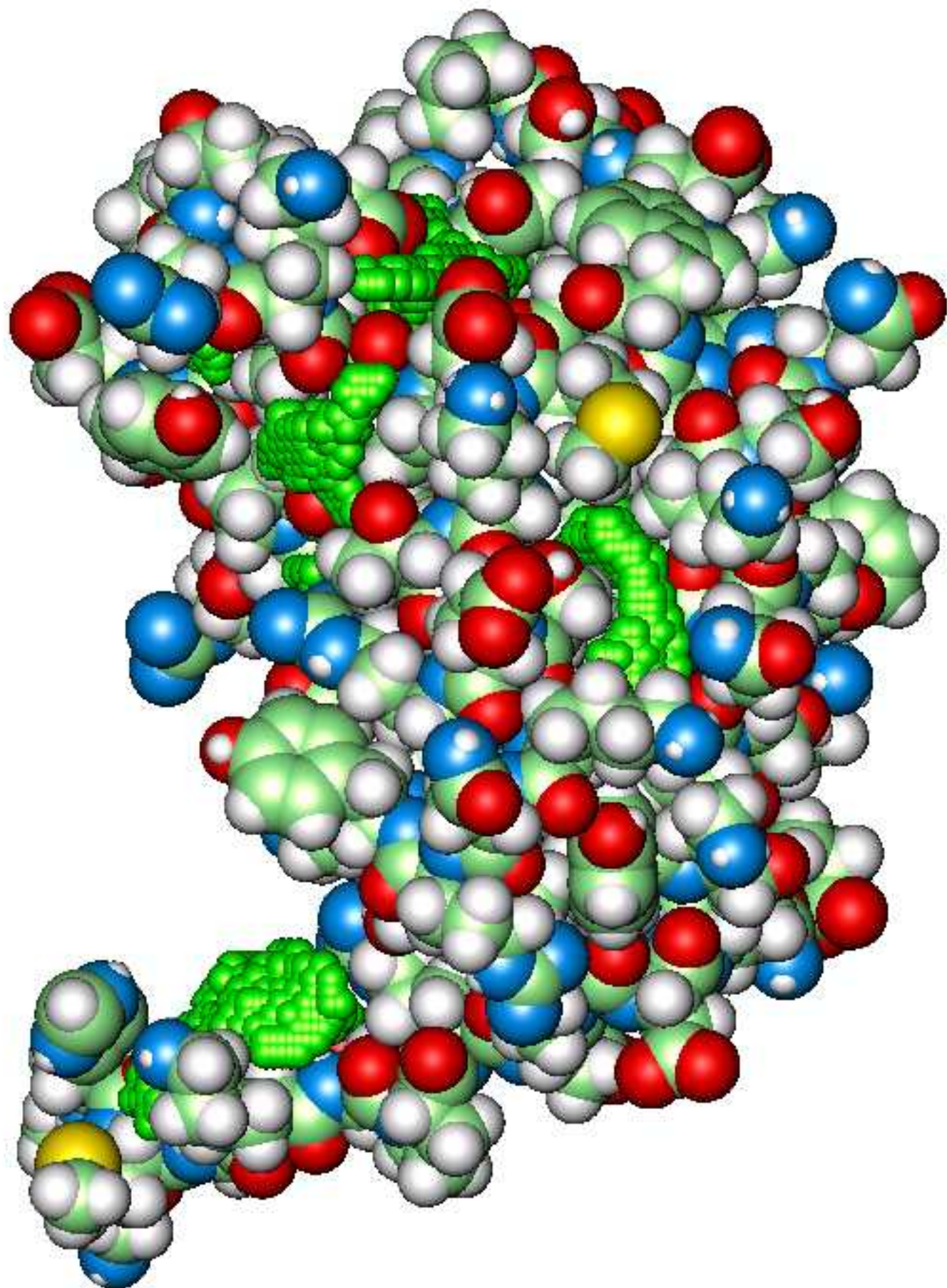


Figure 2



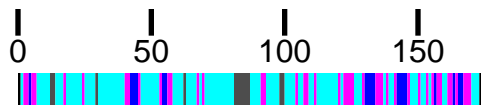
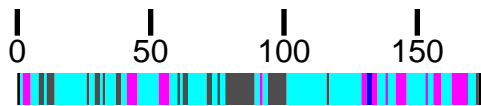
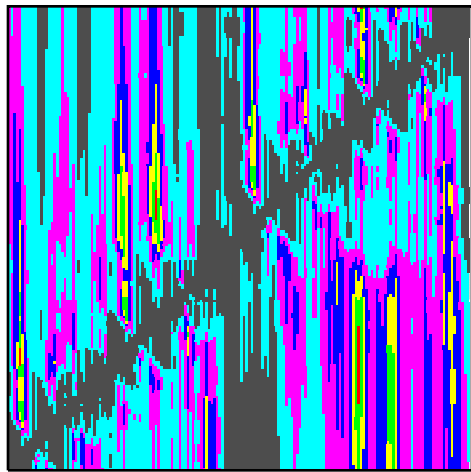
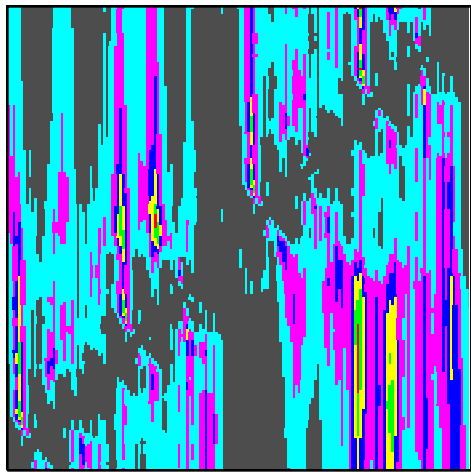






$CV^W$

CV



<1/7: ■ <2/7: ■ <3/7: ■ <4/7: ■ <5/7: ■ <6/7: ■ <7/7: ■

Figure 7

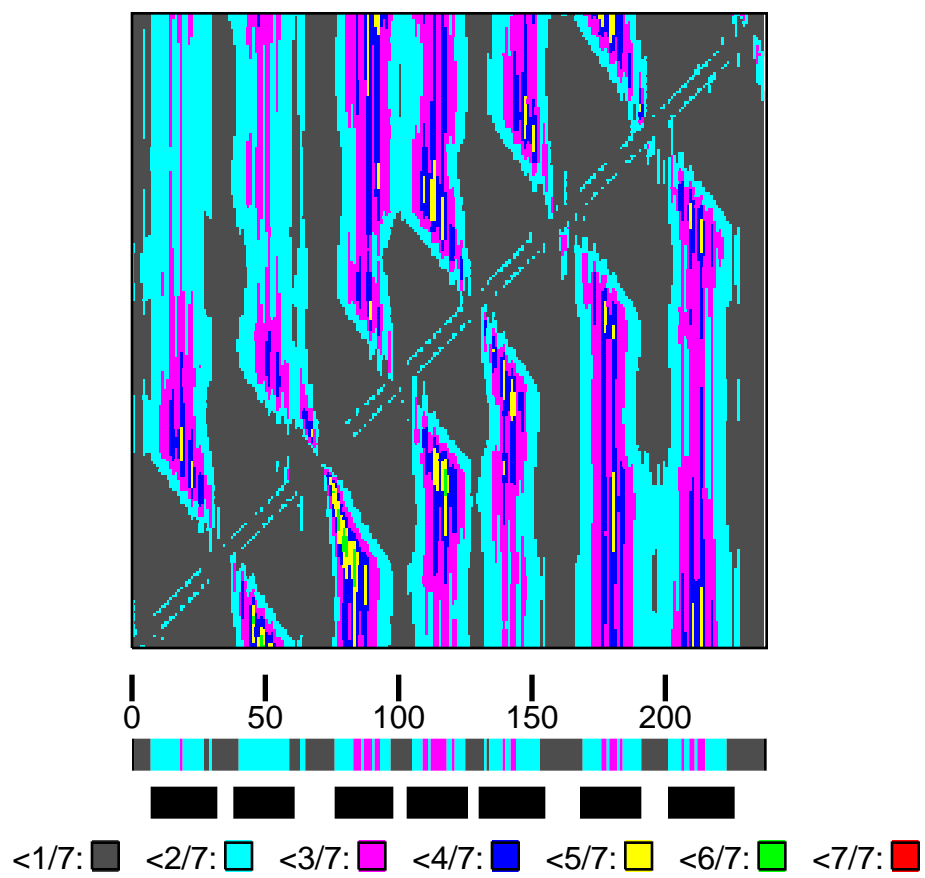


Figure 8