# Towards Optimal Views Of Proteins

preprint of paper to appear in *Bioinformatics*

Oscar Sverud and Robert M. MacCallum
Stockholm Bioinformatics Center
Stockholm University
106 91 Stockholm
Sweden

January 6, 2003

## Abstract

**Motivation:** Graphical representations of proteins in online databases generally give default views orthogonal to the PDB file coordinate system. These views are often uninformative in terms of protein structure and/or function. Here we discuss the development of a simple automatic algorithm to provide a "good" view of a protein domain with respect to its structural features.

**Results:** We used dimension reduction with the preservation of topology (using Kohonen's self organising map) to map 3D carbon alpha coordinates into 2D. The original protein structure was then rotated to the view which corresponded most closely to the 2D mapping. This procedure, which we call OVOP, was evaluated in a public blind trial on the web against random views and a "flattest" view. The OVOP views were consistently rated "better" than the other views by our volunteers.

**Availability:** The source code is available from the OVOP homepage:
`http://www.sbc.su.se/∼oscar/ovop`

## Introduction

When we look at a three-dimensional protein structure in a program such as Rasmol [12], we spend the first few seconds almost instinctively rotating the molecule to a view where we can see "what is going on" in the structure. Unfortunately, the majority of the 16000+ protein structures in the PDB [3] have not been rotated to an informative view. If the structures have been determined by crystallography, the orientation generally corresponds to one of the crystal lattice axes. The result is that web resources such as SCOP, PDBsum and the PDB itself do not always do justice to the beauty of protein structures when default graphical views are given.

The concept of a "good view" is subjective. It depends on what the viewer wants to see; it could be the active site, the sheet topology, or a flexible linker region, for example. However, if we limit the criteria to structural features, i.e. the path of the polypeptide chain through space and the arrangement of secondary structures, we believe that an automatic algorithm can be used to provide good views.

Our approach is very simple: we perform a 3D→2D mapping of carbon alpha coordinates and then rotate the 3D structure to give a 2D view which corresponds best to the 2D mapping. A number of other groups have used 3D→2D dimension reduction for protein structures. Barlow and Richards used the Sammon mapping algorithm to create 2D representations which preserve the major aspects of secondary structure organisation [2]. The current most widespread 2D representation of proteins is the TOPS cartoon [14, 5, 8]. These automatically generated cartoons show very clearly the topological arrangement of the secondary structure elements, which is often difficult to see, even in true 3D. Furthermore, the TOPS descriptions are searchable.

In this work we use Kohonen's self organising map (SOM) for dimension reduction [6]. This neural network related algorithm is randomly initialised and training is unsupervised and competitive. During training, the amount of adaptive learning is slowly decreased as the data organises itself "meaningfully" into the lower dimensional space. Previous experience with SOM mappings of protein structure [10, 13] showed that the important topological relationships are preserved in two dimensions. Figures prepared for that work showed SOM mappings next to standard 2D views of proteins rotated by hand to approximately the same orientation as the mappings. Here we have developed a simple program to perform the mappings and the final rotation automatically. The more difficult task is to measure the quality of the views generated. However, we were fortunate to get a good response to our blind evaluation web questionnaire; independent protein experts consistently rated the SOM-derived view better than the less sophisticated "as flat as possible" and two random control views.

## Methods

### The self organising map

The SOM takes multi-dimensional input data points and maps them to discrete points on a 2D lattice or grid. In this application, the input data points are the $(x, y, z)$ coordinates of the backbone carbon alpha (CA) atoms of each amino acid residue in a protein of known structure. The output grid is a rectangular lattice with the number of nodes approximately equal to the number of residues in the protein. The ratio between the grid's height $(H)$ and width $(W)$ is calculated from the two largest principal components of the input data, since we have found that long thin proteins map better to long thin grids.

Each grid node is represented by a single $(x, y, z)$ coordinate, which is randomly initialised before training commences (within the ranges of the input data). Each training step involves finding for each input data point the closest grid node (having the smallest Euclidean distance between the input coordinate and the grid node coordinate). The coordinates of the "winning" node and its neighbours within a radius, $r$, are then adjusted towards the input data point according to a learning rate, $\alpha$, and a Gaussian neighbourhood function, $N(r)$. All inputs are presented in random order a total of three times, with linearly decreasing $\alpha$ and $r$. The initial values for $\alpha$ and $r$ are 0.05 and $0.5\sqrt{H \times W}$. The result is that each CA atom is mapped to a grid node (Figure 1). Note that two or more data points may map to the same grid node. For the mappings we used the SOM_PAK software available at `http://www.cis.hut.fi/research/som_pak/`.

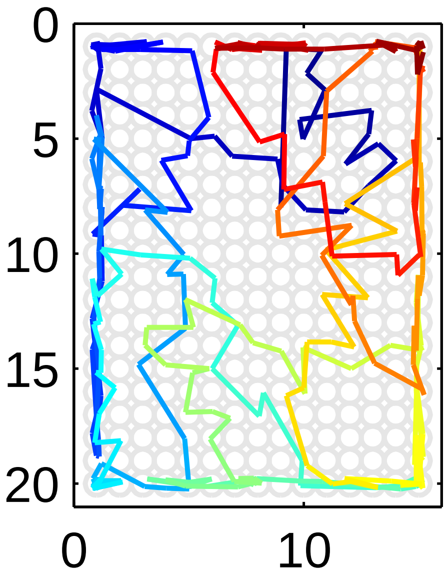


Figure 1: The CA backbone of a TIM barrel protein (PDB: 1tax chain A [9]) mapped with the Self Organizing Map algorithm on to a 20x15 grid. The trace is coloured from blue at the N-terminal through to red at the C-terminal

Because the map vectors are randomly initialised, no two mappings of the same protein are identical. To ensure high quality mappings, we repeatedly sample 20 mappings of a protein until the sampling fails to find a higher quality mapping.

Mapping quality is calculated as the Pearson correlation coefficient between two sets of intramolecular distances. One set consists of the Euclidean distances in 3D space, $d^{3D}_{i,j}$, between all pairs of residues $i$ and $j$. The other set consists of distances in SOM output grid space, $d^{SOM}_{i,j}$, between the same residue pairs. So if residue $i$ maps to SOM grid point $(m_i, n_i)$ and residue $j$ maps to grid point $(m_j, n_j)$, $d^{SOM}_{i,j} = \sqrt{(m_i - m_j)^2 + (n_i - n_j)^2}$. Thus we have the mapping quality measure $r = \text{Pearson}(d^{3D}_{i,j}, d^{SOM}_{i,j})$.

### Rotation to SOM view

The next step is to rotate the protein 3D coordinates to the orientation where the 2D view down the $z$-axis (the standard view in programs like Rasmol) most closely corresponds to the “view” given in the SOM mapping (as in Figure 1). The measure of agreement between the two views is the correlation coefficient, $s = \text{Pearson}(d^{2D}_{i,j}, d^{SOM}_{i,j})$, between the Euclidean distances in 2D view space, $d^{2D}_{i,j}$ (conveniently calculated using just the $x$ and $y$ coordinates of the CA atoms), and SOM space, $d^{SOM}_{i,j}$, as previously defined.

To find the best view (maximal $s$) we could perform an exhaustive search, rotating the protein around both the $x$ and $y$ axes with angles $\phi$ and $\psi$ in, for example, one degree steps. This would however be very time consuming as there are $180^2 = 32400$ different views to compare and the distance correlations are expensive to calculate. Fortunately, initial observations on a small sample of proteins showed that the plots of $s$ against $\phi$ and $\psi$ had a very smooth landscape (data not shown). Assuming that there are not too many examples with multiple maxima, we can use a gradient ascent algorithm to find the best correlated view. The landscape is explored by measuring $s$ at three different rotations separated by 5 degrees. This gives three points $(\phi, \psi, s_1)$, $(\phi, \psi + 5, s_2)$ and $(\phi + 5, \psi, s_3)$, which define a plane. A move is then made (to a new $\phi, \psi$) in a direction perpendicular to the intersection of the plane and the basal plane (i.e. up the slope), with a magnitude proportional to the gradient. When the gradient is less than a small predefined constant, the search is terminated. We will now call this the OVOP view.

## Flat and random views

In effect, the SOM algorithm flattens out the input data onto the 2D grid while trying to retain as many of the original neighbour relations as possible. Therefore it seemed appropriate to test the OVOP view against a simple “flat” view. The flattest view was defined as the maximum (by gradient ascent, see above) of the sum of distances, $\sum d^{2D}_{i,j}$, at a given rotation $(\phi, \psi)$.

Random views were generated by rotating the original PDB coordinates by two random $\phi$ and $\psi$ angles. Our initial assumption was that these views would be uniformly distributed, just like random points on the surface of a sphere. However, one of the referees pointed out that actually our views are biased towards the poles. Basically, the two rotations are not independent of each other, and for certain values of $\phi$, $\psi$ has little effect. It is unfortunately too late to correct this error, but thankfully the bias is not too severe, and original orientations in the PDB are essentially random with respect to the fold topology anyway.

## Blind web trials

A representative set of 1924 protein domains was taken from SCOP [11] release 1.53 for testing the OVOP method. The ASTRAL [4] subset where no pair of sequences has more than 10% sequence identity was used, and further filtered to remove multi-chain domains and small domains with fewer than 50 residues. All SCOP classes were allowed, so some “multi-domain proteins” are included in the set, even though we sometimes refer to them as “domains”.

An email was sent to colleagues (personally known by us) asking them to participate in the testing of OVOP by visiting our web page. The test involves looking at four different Molscript [7] cartoon views of a protein domain chosen at random from the data set. The visitor is asked to rate each view as “Good”, “OK” (default) or “Bad”. There was no option to abstain. The four views are OVOP, flat, and two random views, shown in random order. The views are not labelled and the filenames of the images were loosely encrypted so that the visitor has no way to find out which view is which. The volunteers were asked to rate 20 proteins (80 views) in this way. We intentionally did not give any instructions on *how* to rate the views, except in terms of “Good”, “OK”, or “Bad”. We did, however, suggest that the test would take around 7 minutes. The collection of data has ceased, but the trial can still be performed at `http://www.sbc.su.se/∼oscar/ovop`.

# Results and discussion

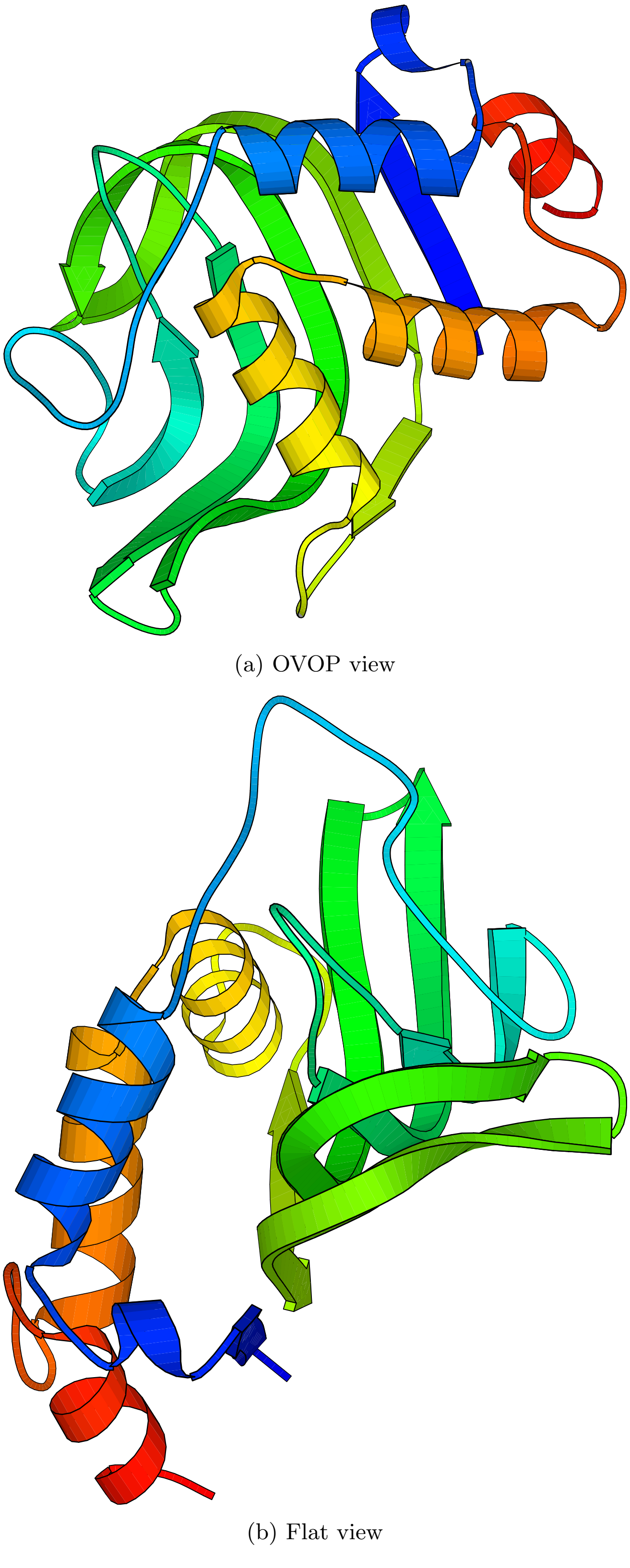
## OVOP views

We have provided just one example of an OVOP view (compared with a flat view for the same protein) in Figure 2. The reader is invited to inspect unlimited OVOP views on the web at `http://www.sbc.su.se/∼oscar/ovop`, where it is possible to see views of domains from a particular SCOP class, fold, superfamily etc. This facility was provided after the completion of the blind web trial (see below), so volunteers were not able to learn what OVOP views looked like.

### Web trial results

During the period 5-14 June 2002, 164 volunteers (many of them members of SBNet, Structural Biology Network, Sweden) evaluated four different views of between 9 and 60 proteins (see Methods). The majority rated 20 or 30 proteins. The total number of evaluations was 3463 (13852 images were viewed and rated). If we apply a simple scoring scheme: views rated "Good", "Bad" and "OK" score +1, −1, and 0, respectively, then it is simple to calculate the mean score for each type of view, OVOP, flat and random 1&2 (Figure 3(a)). The mean score for OVOP is 0.226, reflecting the underlying data: 41.6% of the OVOP views were rated "Good", 18.9% were rated "Bad", and 39.5% "OK". The flat views' mean score is close to zero ("OK on average") and the random views get negative scores, both around -0.2. Using the Student's *t*-test to compare the sampled means of OVOP and flat view scores we find a highly significant difference ($P \ll 0.001$); however a bold assumption is made that our 3-state discrete scores follow a normal distribution, which of course they cannot. Alternatively, the generally close agreement between the data for the two random views (which are methodologically identical) gives a simple visual cue to the underlying error and the adequacy of our sampling.

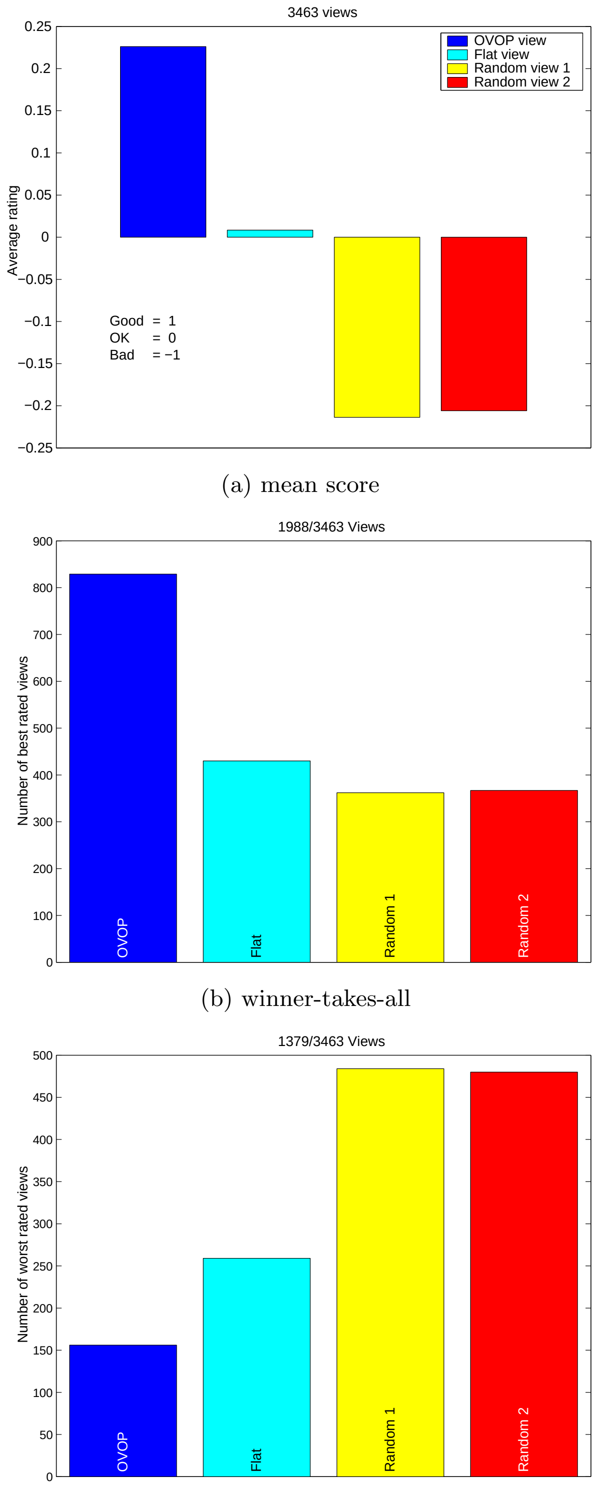
One problem with the mean score approach is that our volunteers have different standards and opinions; they may be more or less likely to rate views as "Bad", for example. Therefore we also processed the ratings using exclusive winner-takes-all and loser-takes-all scoring systems. In winner-takes-all, only the highest rated of the four views (this could be "Good" or "OK") for each protein domain is counted. Where there is a tie, no view type gets a point. The two different random views are treated as separate view types. Of the 1988 proteins with a clear winning view, 829 (41.7%) views were generated by OVOP, around double the number of views generated by the flat or random methods, which perform similarly (Figure 3(b)). The percentage of *all* trials with OVOP the clear winner is, of course, lower (829/3463 = 23.9%). In terms of worst views (Figure 3(c)), OVOP again



(a) OVOP view

(b) Flat view

Figure 2: Example views for heme-binding protein A (PDB code 1b2v [1]). This is a favourable example where the OVOP view is more informative than the flat view. Note that for many proteins the flat view is quite similar to the OVOP view.



(a) mean score

(b) winner-takes-all

(c) loser-takes-all

Figure 3: Summary of web-based human evaluation results for different protein view generation techniques. In (a) the average score (Good scores +1, OK=0, Bad=-1) for each type of view is shown. In (b) only the exclusively highest rated view from the four is counted (winner-takes-all). In (c) only the exclusively lowest rated view is counted (loser-takes-all).

does well, with the fewest (156) worst views. Also note that the difference between flat and random views is quite a lot greater in Figure 3(c) compared to Figure 3(b); the flat view may not be much "more good" than random views, but it is considerably "less bad".

Figure 4 shows the raw and mean score results for the SCOP domains broken down by secondary structural class. The all-$\alpha$ proteins, with their less elaborate topology, are rated the highest across all view types (the only mean which is greater than zero, see the dashed lines in Figure 4). This indicates that, to some extent, "any view will do". Indeed, here we see by far the best performance by the flat view, although OVOP still provides better views. The hardest class to orientate well appears to be the $\alpha/\beta$ domains, where the overall mean and random view means are the most negative. However, it is in this class that OVOP performs the best (only marginally higher than all-$\alpha$). Both OVOP and flat views have lower mean scores in the all-$\beta$ class than in the other classes, although OVOP still gives a positive mean score. Clearly, the all-$\beta$ class presents the greatest challenge to the OVOP method, and we discuss this in more detail later.

In the Methods section we mentioned that some of the SCOP domains are in fact multi-domain. An obvious problem arises when the best view for one domain in a multi-domain protein does not correspond to the best view for other domain(s). The SOM algorithm will fit sub-domains into the grid in different orientations as necessary. The compromise comes when we rotate the 3D coordinates to the best view. The same problem exists for large structures where there are more secondary structure elements to consider. We also note that many proteins in the first four classes of SCOP are in fact not fully split into domains; this is particularly true for the immunoglobulin-like folds. Therefore we looked at the effect of length on the perceived quality of protein views. Figure 5 clearly shows decreasing quality of flat and random views with increasing length of the proteins. The quality of OVOP views does not directly follow this trend. The quality drops a little for the largest group of proteins (> 300 residues), but we surprisingly see the poorest performance for the smallest protein domains (50-100 residues). We expected perhaps that the small domains would attract an excess of "OK" ratings, but in fact the lower mean score is the result of more "Bad" rat-

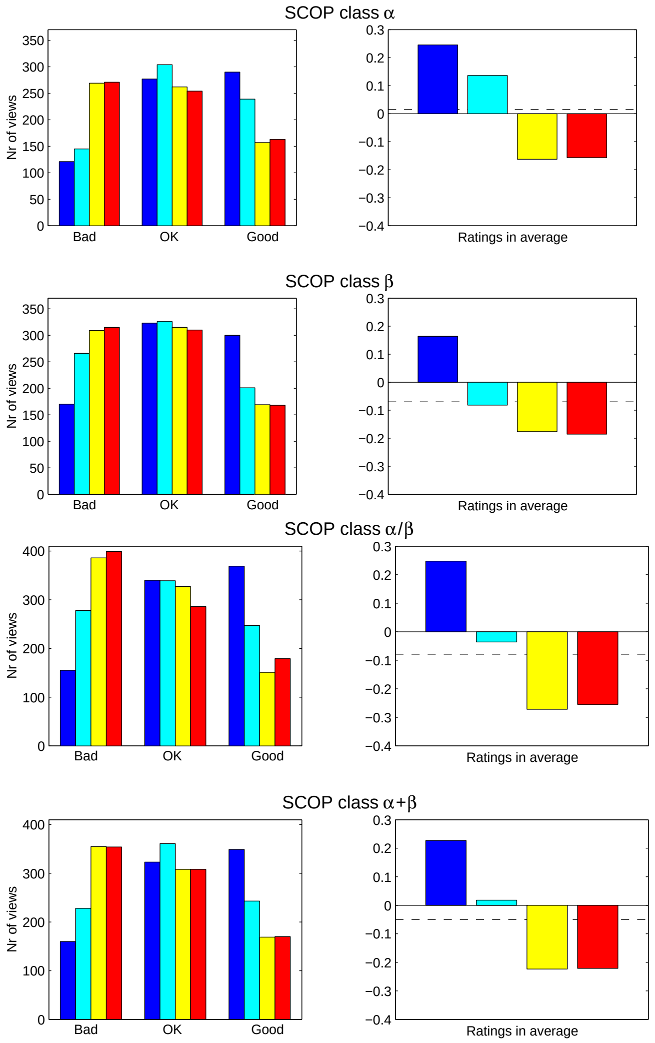


Figure 4: View evaluation results broken down for the four main SCOP classes. Left panels: raw results. Right panels: mean score summary (as in Figure 3(a)), the horizontal dashed lines indicate the mean score for all view types combined. The bar colours indicating the view types are the same as in Figure 3.

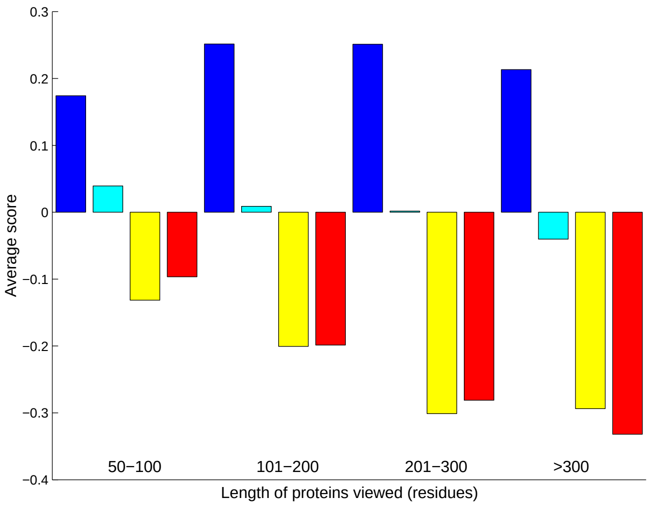


Figure 5: Mean scores for different views of proteins in different length classes. For legend, see previous figures.

ings than expected (data not shown).

Finally we show the relative performance of OVOP for five of the most heavily populated SCOP folds, which give us enough data to analyse (Figure 6). The immunoglobulin-like $\beta$-sandwich (SCOP b.1) is the second most commonly occurring SCOP fold in multicellular organisms (first is the "classic zinc-finger C2H2" fold, data from the 3D-Genomics web resource, `http://www.sbg.bio.ic.ac.uk/3dgenomics`), and is used widely in extracellular molecules involved in cell to cell contacts and communication. It is also the fold, of these five, where OVOP is least-good and the random orientation is least-bad. This is probably because $\beta$-sandwiches are quite difficult to view, even by hand; views orthogonal to the sheets (generally favoured by OVOP) tend to obscure one sheet behind the other, and views from the side fail to show the strand connections clearly.

In contrast to the immunoglobulins, the TIM-barrel fold (SCOP c.1) has the best OVOP views and the worst random and flat views. In the most common all-$\alpha$ fold, LEM/SAP HeH motif (SCOP a.4), the flat views are rated fractionally better than the OVOP views, which follows the trend seen in Figure 4. The wide variation in OVOP's behaviour is not surprising since the protein universe is large and diverse. When OVOP does not perform so well for a large group of proteins, like the immunoglobulins, it would be quite possible to take a different approach: where one domain is rotated to the agreed expert view from the literature (these exist for the immunoglobulin and globin folds, for example) and other domains are structurally aligned to the same view.

## Future developments

We received many useful comments from participants in the web trial, and summarise them here:

- The images are too small for large proteins, and for all proteins the scaling is not consistent.
- The participant did not know the SCOP code of the domain being viewed, and so could not rotate the molecule on his/her own computer. (However, we didn't want people to do this, since it would take longer and less people would complete the survey.)
- We do not take into account the participant's reaction to the aesthetic qualities of protein structures (perhaps this explains the high score of the TIM-barrel fold).
- The participant's criteria for "Good/OK/Bad" almost certainly change *during* the survey.

Both the OVOP method and the evaluation methods could be improved. Since the evaluation only gives a relative rating of the methods, it is important to work on the control methods too. For example the "flat" method could be enhanced to give a flat (or slightly inclined) view of the largest sheet (if present). Also, views could be rotated around the $z$-axis to a common orientation of the first strand or helix.

The reader may have already noted that the "front" and "back" views of the protein are identical with respect to the distance correlation calculations. The OVOP view is therefore randomly oriented in this respect. Future versions should attempt to determine which of these two views gives most information (for example, with the fewest obscured secondary structures). Established computer vision techniques may be of some help here and also to find more specialised views. OVOP has been developed primarily to show the entire fold of the protein or domain in a good orientation. However it could be adapted to prioritise sheets or helices or some other region of interest specified by the user or another program.

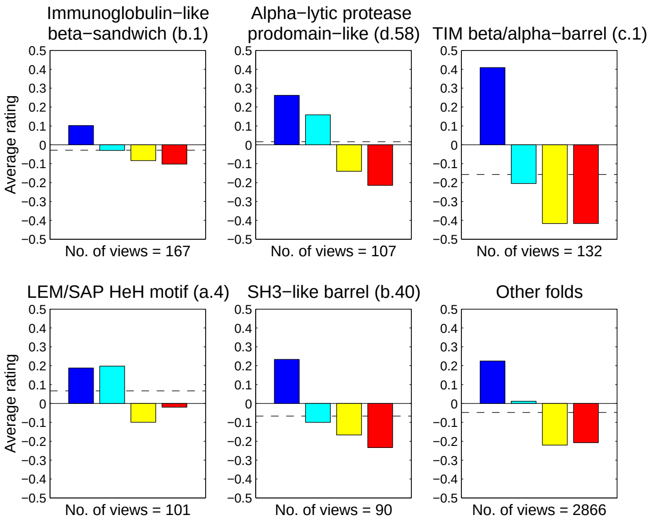


Figure 6: Mean score summary for views of five common folds. For legend, see previous figures.

A major hindrance to further development is "volunteer fatigue"; one has limited opportunities to ask for help from the community in testing new variants of the view generation algorithms. However, based on the results presented here, we feel that the current version of OVOP, which is freely available for non-commercial use, is ready for use in any application wherever arbitrary views are currently given.

## Acknowledgements

We are extremely grateful for the help of all the participants in the web trial, who are unfortunately too numerous to list here. We thank Per Kraulis, Arne Elofsson and Alex May for helpful discussion, and the anonymous referees for their valuable input. The authors and Stockholm Bioinformatics Center are supported by the Swedish Foundation for Strategic Research.

## References

[1] P. Arnoux, R. Haser, N. Izadi, A. Lecroisey, M. Delepierre, C. Wandersman, and M. Czjzek. The crystal structure of HasA, a hemophore secreted by Serratia marcescens. *Nature: Struct. Biol.*, 6(6):516–520, Jun 1999.

[2] T. W. Barlow and W. G. Richards. A novel representation of protein-structure. *J. Mol. Graph.*, 13:373, 1995.

[3] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nuc. Ac. Res.*, 28:235–242, 2000.

[4] S. E. Brenner, P. Koehl, and M. Levitt. The ASTRAL compendium for protein structure and sequence analysis. *Nuc. Ac. Res.*, 28(1):254–256, 2000.

[5] T. P. Flores, D. S. Moss, and J. M. Thornton. An algorithm for automatically generating protein topology cartoons. *Protein Eng.*, 7:31–37, 1994.

[6] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69, 1982.

[7] P. J. Kraulis. Molscript — a program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.*, 24:946–950, 1991.

[8] M. Levitt and C. Chothia. Structural patterns in globular proteins. *Nature*, 261:552–557, 1976.

[9] L. Lo Leggio, S. Kalogiannis, M. K. Bhat, and R. W. Pickersgill. High resolution structure and sequence of T. aurantiacus xylanase I: implications for the evolution of thermostability in family 10 xylanases and enzymes with (beta)alpha-barrel architecture. *Proteins: Struct., Funct., Genet.*, 36(3):295–306, 1999.

[10] R. M. MacCallum. *Computational analysis of protein sequence and structure.* PhD thesis, University of London, http://www.sbc.su.se/∼maccallr/thesis, 1997.

[11] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP — a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 247:536–540, 1995.

[12] R. A. Sayle and E. J. Milnerwhite. Rasmol — biomolecular graphics for all. *Trends Biochem. Sci.*, 20:374–376, 1995.

[13] O. Sverud. *Self Organizing Maps of Protein Structure.* Masters thesis, Royal Institute of Technology (KTH), Stockholm, http://www.sbc.su.se/∼oscar/downloads/thesis/thesis.pdf, 2002.

[14] D. R. Westhead, T. W. Slidel, T. P. Flores, and J. M. Thornton. Protein structural topology: Automated analysis and diagrammatic representation. *Prot. Sci.*, 8(4):897–904, 1999.