Representation of Protein Secondary Structure Using Bond-Orientational Order Parameters

Cem Meydan and Osman Ugur Sezerman

Sabanci University, Biological Sciences & Bioengineering Dept., Istanbul, Turkey
{cemmeydan,ugur}@sabanciuniv.edu

Abstract. Structural studies of proteins for motif mining and other pattern recognition techniques require the abstraction of the structure into simpler elements for robust matching. In this study, we propose the use of bond-orientational order parameters, a well-established metric usually employed to compare atom packing in crystals and liquids. Creating a vector of orientational order parameters of residue centers in a sliding window fashion provides us with a descriptor of local structure and connectivity around each residue that is easy to calculate and compare. To test whether this representation is feasible and applicable to protein structures, we tried to predict the secondary structure of protein segments from those descriptors, resulting in 0.99 AUC (area under the ROC curve). Clustering those descriptors to 6 clusters also yield 0.93 AUC, showing that these descriptors can be used to capture and distinguish local structural information.

Keywords: bond-orientational order, secondary structure, machine learning, structural alphabet.

1 Introduction

In analysis protein structures, different models of representations on various levels of structural details are used. From coarse-grained to all-atom models, simplified lattice to continuous representations, each model can be used in different areas of research.

The need for abstraction in computational methods (such as structure search and comparison, fold matching, structural motif mining and other areas of pattern recognition) is especially high. The very high amount of data and precision in the 3D coordinates makes computational analysis very complex and very rigid in its applicability. Simplified models capture relevant information and hide unimportant details through abstraction, conferring the ability to group complex 3D information into manageable clusters that can be searched for, compared and "learned" by machine-learning algorithms in a flexible fashion.

The most common simplified representation of the protein states are the secondary structural assignments to the coordinates, which can be overlaid onto the sequence to create a 1D representation.

There have been other studies with aims to create local structural alphabets to represent the structure as a 1D sequence of structural blocks [1]. A structural alphabet

T. Shibuya et al. (Eds.): PRIB 2012, LNBI 7632, pp. 188–197, 2012.

is defined as a set of small prototypes that can approximate each part of the backbone. Creating such an alphabet requires the identification of a set of recurrent blocks that can identify all possible backbone conformations. A commonly used structural alphabet is PB [2], which uses the dihedral angles of the backbone structure in a sliding window to match the segment to one of 16 pre-defined blocks.

Another common approach for structure abstraction is to convert the protein structure into a graph from distance or contact maps. In this representation, each residue is coarse-grained into one center node that is connected to other nodes on the graph on the basis of distance (or other criteria). This allows each aminoacid to be represented with its contacts and the topology of the network around it. Representing the structure as a graph allows for sub-graph matching to find reoccurring common motifs in a data set [3], use of elastic network models for normal mode analysis [4] and other algorithms that can employ the graph theoretical properties.

The problem with different representation schemas is the amount of information lost to the abstraction. In case of secondary structure, representing the structure with two states (α-helix and β-sheet) causes the diversity of helices and sheets to be lost, as α-helices are frequently curved (58%) or kinked (17%) [5]. Use of local structural alphabets can capture this information; however as the name implies, the non-local neighbor information of the protein structure is missing. Graph based methods can capture both local and global information from the graph topology. However, since the 3D coordinates of the contacts are reduced to only edge weights, direction and the topology of the structure around each residue is lost.

To approximate both the local structural information with a relatively high degree of certainty and the non-backbone neighbor information and directionality of the contacts with a single model, we propose the use of bond-orientational order parameters. Bond-orientational order is a well-established metric that is used in analysis and comparison of the crystal structures packing of atoms [6]. Due to the use of spherical harmonics, they can capture the directional information around each residue, and since they are invariant of the rotations of the reference frame, matching two structures require only the comparison of numbers, instead of the more computationally costly and problem-prone structural alignment methods.

As a first step, we wanted to test whether the number and placement (angle and distance) of neighboring atoms around each residue show a repeating pattern in average protein structures. If there is such a pattern, we can use the protein descriptors to approximate the local structure around a center point. To test the feasibility of representing the protein structure with such orientational order descriptors, we tried to use those descriptors to capture and differentiate the secondary structural elements from each other. Recognizing and assigning secondary structures to atomic coordinates is a complex task [7] and require the ability to recognize both the local structure (for helices) and contact information (between β strands). If orientational order descriptors can predict secondary structural elements, it shows that they capture the necessary information and can be evaluated further for more complex motif discovery purposes.

2 Methods

2.1 Bond-Orientational Order

The bond-orientational order parameter is previously described by Steinhardt et al. [6] in the study of packed spheres. It has also been employed in the analysis of protein structures by means of local connectivity around each residue [8]. The bond-orientational order parameters are given as:

$$\bar{Q}_{lm}(i) = \frac{1}{N_b(i)} \sum_{n=1}^{N_b(i)} Y_{lm}[\theta(\vec{r}_n - \vec{r}_i), \phi(\vec{r}_n - \vec{r}_i)] \tag{1}$$

$$Q_l(i) = \left(\frac{4\pi}{2l+1} \sum_{m=-l}^{l} |\bar{Q}_{lm}(i)|^2 \right)^{1/2} \tag{2}$$

$$W_l(i) = \sum_{\substack{m_1,m_2,m_3, \\ m_1+m_2+m_3=0}} \begin{bmatrix} l & l & l \\ m_1 & m_2 & m_3 \end{bmatrix} \times \bar{Q}_{lm_1} \bar{Q}_{lm_2} \bar{Q}_{lm_3} \tag{3}$$

where l is the bond orientational parameter. $\vec{r}_n$ denotes the position vector of the nth residue. $(\vec{r}_n - \vec{r}_i)$ is the bond vector from residue i to n, and θ, ϕ are the polar angles of this bond, measured with respect to an arbitrary reference frame. $Y_{lm}[\theta(\vec{r}_n - \vec{r}_i), \phi(\vec{r}_n - \vec{r}_i)]$ are Laplace's spherical harmonic functions [9] for the given angles. $N_b(i)$ is the total number of contacts of i that are below a given cutoff distance. The coefficients shown as a matrix in Equation 3 are the Wigner-3-j symbols [10].

While the spherical harmonics of the bonds for a given l can change drastically by rotating the coordinate system, combining the Q_{lm} values into a quadratic invariant Q_l (Equation 2) and third-order invariant W_l (Equation 3) will result in a rotationally invariant parameter. These order parameters are invariant under reorientations of the external coordinate system. For l=2n, spherical harmonics are also invariant under inversion and therefore independent of reference frame.

In research of the crystal packing, most commonly used parameter is the Q_6 [6, 11, 12] as l=6 is the smallest value of l that can capture both cubic (simple, face centered, and body centered) and icosahedral orders (whereas Q_4 will miss icosahedral and Q_2 will miss both) [8].

2.2 Dataset

For experimentation, a total of 120 protein structures were collected from the Protein Data Bank [13]. Protein structures belonging to different SCOP [14] classes and folds were selected for more even representation of different folds in the dataset. On top of those, the benchmark set of non-homologous (<30% sequence identity) PDB proteins of Zhang et al. [15] were also added to the final dataset.

For each residue, Q_l (Equation 2) and W_l (Equation 3) values are calculated from the contacts of that residue, where contact is defined as residues with distance between the Cα atoms that is less than a predefined cutoff threshold. During the calculation, different cutoff distances and l values were tried. Resulting Q_l and W_l values were merged in a feature vector by using sliding window on the backbone.

The secondary structure of each protein was calculated using STRIDE [16]. The secondary structure values of the windows were assigned as a class value on the basis of occurring in the majority of the segment (>60%) in a continuous fashion in the sliding window. The transition regions between different secondary structures that contain two or more different secondary structure classes in the protein segment were removed from the dataset since there is no clear secondary structure to be used in learning and prediction. After those removals, extracting the features from the 120 proteins (using a window size of 5) results in 15273 rows (protein segments) in the final dataset.

2.3 Secondary Structure Prediction

Secondary structures assigned to protein segments by STRIDE [16] are represented in a 3-class and 7-class fashion. The 7 classes are α helix, 3_{10} helix, π helix, β-sheet, coil, turn and bridge. Those 7 classes were simplified to 3 classes as "Helix", "Sheet" and "Loop". The final dataset was created using both 3-class and 7-class representations. However, in the resulting dataset the classes bridge, 3_{10} helix and π helix had only few copies, as either they are uncommon or are rarely found consecutively. Also, STRIDE is believed to underpredict π helices [17], possibly lowering their count even further. Due to very low sample size, 3_{10} helix and π helix classes were merged with the alpha helix class, and bridge regions were removed completely, resulting in a 4-class (helix, sheet, coil, turn) data.

In the feature vector, Q_l values always result in a value between 0 and 1, while W_l values can take arbitrary values. To overcome this, W_l values were normalized to the [0-1] range before the prediction.

Using the calculated Q_l and normalized W_l values from the sliding windows as the feature vector, and assigned secondary structure as the class value (for both 3-class and 4-class), a classification was performed using the SVM implementation libsvm [18] inside the Orange data mining software [19].

Optimization of the window size, l-values and the cutoff distance was carried out on a smaller independent set consisting of 15 proteins. The optimal results were obtained using a cutoff of 7 Å in conjunction with l=2 to 10, with a window size of 5.

Training and prediction was done on separate datasets, created from independent proteins (i.e. no protein segment was predicted with a classifier that was trained with a segment belonging to the same protein). The data was split in a 50-50% fashion (of the PDBs) to create the training and the testing sets.

3 Results

3.1 Prediction Results

The accuracy and the AUC (area under the receiver-operating-characteristic curve) of the predictions of the test set are given in Table 1. Accuracy of the prediction is 92.3% and the AUC is 0.993. AUC gives the probability that a randomly selected positive instance will score higher than a random negative instance, and is a more robust performance measure than accuracy itself [20].

Looking at the confidence table, helices (sensitivity of 0.99) can be represented exceptionally well by the bond-orientational order parameters, followed by sheet structures (sensitivity of 0.91). In 4-class representation, coils and turns have lower sensitivity (respectively 0.71 and 0.75). However, as can be expected, they are more likely to be mistaken as each other than a sheet or helix. In 3-class representation, assigning the class value of "loop-region" to coils and turns will result in a significantly higher sensitivity of 0.87.

Table 1. Area under the ROC curve, accuracy and confusion matrix of the test set predictions. In the confusion matrix, number of predicted instances and ratio of the correct predictions are given.The last row (C+T) represents Coil and Turns being classified as Loop-region in the 3-class prediction.

AUC	Accuracy
0.993	**92.3%**

		Predicted				
		Helix	**Sheet**	**Coil**	**Turn**	**Sensitivity**
Actual	**Helix**	3768 (98.9%)	12 (0.3%)	0 (0.0%)	26 (0.7%)	0.990
	Sheet	3 (0.2%)	1199 (90.70%)	12 (0.9%)	105 (7.9%)	0.907
	Coil	3 (0.7%)	27 (6.1%)	316 (71.0%)	99 (22.2%)	0.710
	Turn	71 (10.1%)	43 (6.1%)	48 (6.9%)	527 (75.3%)	0.753
	C+T	74 (6.5%)	70 (6.1%)	990 (87.3%)		0.873

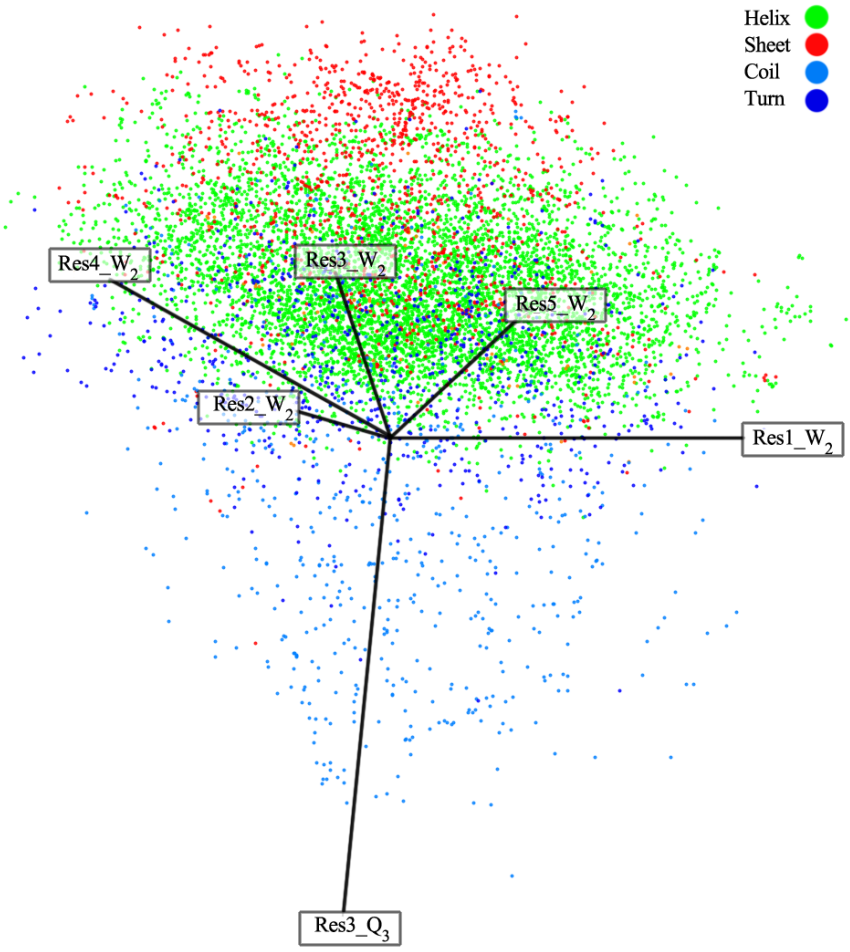

Fig. 1. Distribution of the class values with respect to different features in 2D linear projection. ResX_Y represents the feature Y of the residue X (out of 5) in the proteing segment.

3.2 Feature Analysis and Clustering

Due to the very high accuracy and AUC values, we investigated whether the high accuracy was because of the high predictive performance of SVM due to the use of non-linear kernels, or whether the accuracy could be replicated with a simple, human-understandable method.

We first investigated the effects of different Q_l and W_l features for each residue in the segment to the corresponding secondary structure. To see the importance of each feature and a visual representation of their relationship with samples, we created a 2D linear projection [21] of the data using 6 features, selected by running the VizRank heuristic [22] for 2000 generations on the training set. The rotation of the axes and the final projection was optimized using the FreeViz algorithm [23] to optimize

separation of data points. The result is given in Figure 1. From the perspective of the Q3 and W3 parameters, sheets and coils form the opposing ends of the spectrum. Notice that the classes show a non-perfect but distinct separation even on a linear projection.

To further investigate the quantitative importance of each feature to the prediction, we looked at the information gain and the linear SVM weight of the features. The features that have the highest information gain are the Q_3 and Q_4 values for the middle 3 residues of the window of 5 aminoacids. When ranked by their SVM weights, Q_9 values of the middle 3 residues were also selected as well as the Q_4 values. Not surprisingly, the center portion of the window was ranked higher than the boundary portions. No W_1 values were selected as informative. We can conclude that Q_3, Q_4 and Q_9 are the most important features for classification, since they were all selected at least 3 times for that center portion without exception.

Using the top 6 features from the SVM weights (Q_4 and Q_9 for the 3 center residues of each window), we performed unsupervised k-means clustering on the dataset. The distance between each row was calculated as the distance between their vectors. Euclidean, Manhattan, Hamming distances and Pearson and Spearman correlation values were tried during the clustering. The optimal distance measure was found to be the Manhattan distance. Results for clustering with k=6 in k-means algorithm are given in Figure 2 and Table 2. Figure 2 shows the frequency of the secondary structural elements in the resulting clusters, and Table 2 gives the clustering accuracy and relative assignments of each class to each cluster.

As we can see, even after discretizing the feature vectors to only 6 clusters with an unsupervised method, the clustering has 84.6% accuracy and 0.932 AUC. The clusters show relatively high sensitivity. That is, clusters 1,2 and 3 can represent helix structures with high certainty, cluster 4 is mostly sheet structures and the cluster 5, 6 is commonly loop regions, with most of the errors are due to misclassifying “Turns” as “Coils” and vice versa.

4 Discussion

In our study, we tested the feasibility of using bond-orientational order parameters as descriptors of protein structure in predicting secondary structure from the coordinates Cα atoms. This resulted in 92.3% accuracy and 0.993 AUC. The helices can be predicted at ~99% sensitivity. Since helices are formed by local interactions that are established within the close vicinity of each amino acid, we can conclude that this structure can easily be captured by the orientational order parameters.

While helices can be predicted quite easily using backbone dihedral angles, this is not the case for sheet structures due to non-local, long range interactions. We show that orientational order parameters can capture the representation of β-sheets equally well (91% sensitivity) since strands stand parallel to each other to form the sheets. There is less information coming from the sequentially adjacent residues forming the sheet in comparison to helices (which makes it difficult to predict them in secondary structure prediction algorithms) but the orientational order descriptors can still capture

Table 2. Relative assigment of each class to the clusters. Cluster representations show which class is more likely to be in that cluster.

	# Helix	# Sheet	# Turn	# Coil	Representation
Cluster 1	98.3%	1.1%	0.5%	0.2%	Helix
Cluster 2	90.4%	4.1%	3.8%	1.7%	Helix
Cluster 3	90.1%	5.1%	3.1%	1.7%	Helix
Cluster 4	19.7%	68.9%	8.1%	3.3%	Sheet
Cluster 5	5.8%	20.9%	47.3%	26.0%	Loop region ~ Turn
Cluster 6	0.0%	0.7%	35.9%	63.4%	Loop region ~ Coil

Clustering Accuracy	84.6%
AUC	0.932

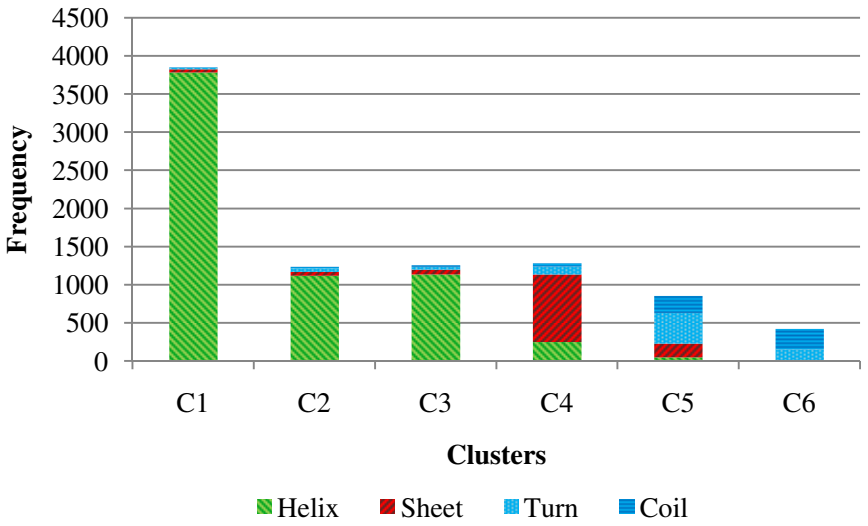

Fig. 2. The number of elements in each cluster by their secondary structural elements

the necessary local and neighbor information. Addition of orientational order parameters with higher cutoff distance values may help in this regard.

Turns and coils are more difficult to predict in comparison to helices and sheets, (75% and 71% sensitivity respectively). This is expected as they are short, can be

found in different local environments (i.e. buried in the core or exposed to water) and lack a rigid structure. Turns are easier to predict than random coils since they are more structured and may have conserved hydrogen bonds between the backbone residues. Some coil structures can be mistakenly classified as turns (22.2%) but the rate of misclassification of turns as coils is not as high (6.9%).

While the continuous features are shown to be enough to capture secondary structure, we also investigated the applicability of comparing two orientational order feature vectors to evaluate structural similarity (i.e. whether a vector can be assigned to a class based on just a distance value and not by a complex rule learned by the SVM). By using an unsupervised clustering method with a simple Manhattan distance metric, we have obtained 6 clusters that correctly predict the secondary structure with 84.6% accuracy and 0.932 AUC, showing that similar structures definitely have similar vector characteristics, which is very important for use in structural alphabets. We can also see this effect in Figure 1; the classes have distinctive characteristics in their features that can be recognized even on a linear projection with few features.

We also looked at the relative importance of each feature in the descriptor vector. Q_3, Q_4 and Q_9 seem to be the most important features in prediction of the secondary structure elements, but a more through experimentation is needed.

We conclude that there is very strong potential application of orientational order parameters, especially in establishment of a new structural alphabet that takes local backbone structure as well as contact information from the neighboring regions into account. Such an alphabet can be exploited to identify structural motifs in a protein family that cannot be captured with other methods.

References

1. Joseph, A.P., Agarwal, G., Mahajan, S., Gelly, J.C., Swapna, L.S., Offmann, B., Cadet, F., Bornot, A., Tyagi, M., Valadie, H., Schneider, B., Etchebest, C., Srinivasan, N., De Brevern, A.G.: A short survey on protein blocks. Biophys. Rev. 2, 137–147 (2010)
2. de Brevern, A.G., Etchebest, C., Hazout, S.: Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. Proteins 41, 271–287 (2000)
3. Grindley, H.M., Artymiuk, P.J., Rice, D.W., Willett, P.: Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. J. Mol. Biol. 229, 707–721 (1993)
4. Atilgan, A.R., Durell, S.R., Jernigan, R.L., Demirel, M.C., Keskin, O., Bahar, I.: Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys J. 80, 505–515 (2001)
5. Martin, J., Letellier, G., Marin, A., Taly, J.F., de Brevern, A.G., Gibrat, J.F.: Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. BMC Struct. Biol. 5, 17 (2005)
6. Steinhardt, P.J., Nelson, D.R., Ronchetti, M.: Bond-Orientational Order in Liquids and Glasses. Phys. Rev. B 28, 784–805 (1983)
7. Offmann, B., Tyagi, M., de Brevern, A.G.: Local protein structures. Curr. Bioinform. 2, 165–202 (2007)
8. Atilgan, C., Okan, O.B., Atilgan, A.R.: How orientational order governs collectivity of folded proteins. Proteins 78, 3363–3375 (2010)

9. Sternberg, W.J., Smith, T.L.: The theory of potential and spherical harmonics. Univ. of Toronto Press, Toronto (1946)
10. Landau, L.D., Lifshitz, E.M.: Quantum mechanics: non-relativistic theory. Pergamon Press; sole distributors in the U.S.A., Addison-Wesley Pub. Co., Reading, Mass., Oxford, New York (1965)
11. Truskett, T.M., Torquato, S., Debenedetti, P.G.: Towards a quantification of disorder in materials: distinguishing equilibrium and glassy sphere packings. Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics 62, 993–1001 (2000)
12. Torquato, S.: Random heterogeneous materials: microstructure and macroscopic properties. Springer, New York (2002)
13. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The Protein Data Bank. Nucleic Acids Res. 28, 235–242 (2000)
14. Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C.: SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536–540 (1995)
15. Zhang, Y., Skolnick, J.: TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 33, 2302–2309 (2005)
16. Frishman, D., Argos, P.: Knowledge-based protein secondary structure assignment. Proteins 23, 566–579 (1995)
17. Fodje, M.N., Al-Karadaghi, S.: Occurrence, conformational features and amino acid propensities for the pi-helix. Protein Eng. 15, 353–358 (2002)
18. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines (2001)
19. Demšar, J., Zupan, B., Leban, G., Curk, T.: Orange: From Experimental Machine Learning to Interactive Data Mining. In: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) PKDD 2004. LNCS (LNAI), vol. 3202, pp. 537–539. Springer, Heidelberg (2004)
20. Hanley, J.A., McNeil, B.J.: A method of comparing the areas under receiver operat-ing characteristic curves derived from the same cases. Radiology 148, 839–843 (1983)
21. Koren, Y., Carmel, L.: Visualization of labeled data using linear transformations. In: In-fovis 2002: IEEE Symposium on Information Visualization 2003, Proceedings, pp. 121–128, 248 (2003)
22. Leban, G., Zupan, B., Vidmar, G., Bratko, I.: VizRank: Data visualization guided by machine learning. Data Mining and Knowledge Discovery 13, 119–136 (2006)
23. Demsar, J., Leban, G., Zupan, B.: FreeViz–an intelligent multivariate visualization approach to explorative analysis of biomedical data. J. Biomed. Inform. 40, 661–671 (2007)