A Supervised Approach to 3D Structural Classification of Proteins

Virginio Cantoni[1], Alessio Ferone[2],
Alfredo Petrosino[2], and Gabriella Sanniti di Baja[3]

[1] University of Pavia, Department of Electrical and Computer Engineering,
Via A. Ferrata, 1, 27100, Pavia, Italy
virginio.cantoni@unipv.it
[2] University of Naples Parthenope, Department of Applied Science, Centro
Direzionale Napoli - Isola C4, 80143, Napoli, Italy
{alessio.ferone,alfredo.petrosino}@uniparthenope.it
[3] Institute of Cybernetics "E. Caianiello" - CNR - Naples - Italy
g.sannitidibaja@cib.na.cnr.it

Abstract. Three dimensional protein structures determine the function of a protein within a cell. Classification of 3D structure of proteins is therefore crucial to inferring protein functional information as well as the evolution of interactions between proteins. In this paper we propose to employ a recently presented structural representation of the proteins and exploit the learning capabilities of the graph neural network model to perform the classification task.

Keywords: Concavity Tree, Graph Neural Network, Structural Classification of Proteins, Protein Function.

1 Introduction

There are currently more than 90,000 experimentally determined three dimensional (3D) structures of protein deposited in the Protein Data Bank (PDB) [25]. 3D structures of a protein are determined by the amino acid sequence and define the protein functions. The study of structural building blocks is very important in order to study the evolution and the functional annotation, and has yielded many methods for their identification and classification in classes of known structure.

There are several commonly used DBs for structural classification of proteins, such as Structural Classification Of Proteins (SCOP) [22] and Class Architecture Topology and Homologous super families (CATH) [23]. CATH and SCOP are primary and secondary structure based classifications which rely on experts to manually check the classifications. Such classifications organize protein structures into families. Another well-known DB is Families of Structurally Similar Proteins (FSSP), which is purely automatic [11].

Most of the existing classification methods are based on shapes, by means of suitable distances, such as RMSG [2]. Consequently, they have to deal with tens of thousands of structures for each protein. These approaches consume significant computation space. Machine learning methods to cluster and classify

A. Petrosino, L. Maddalena, P. Pala (Eds.): ICIAP 2013 Workshops, LNCS 8158, pp. 326–335, 2013.

protein structures have recently become a very active area of research. Statistical methods include Shape Histograms proposed by Anskerst [16], Shape Distribution by Osada [24], meanwhile Ohbuchi put forward Shape Function and eigen-CSS proposed by Mark. In [17], authors use a statistical method as feature vectors to classify the protein structures. Since the statistics methods are based on overall feature extractions, the adoption statistical methods has a high robustness of the boundary noise as well as good performance for rough classification. In particular artificial neural networks (ANN) have been employed to develop a comprehensive view of protein structures to infer protein functional information.

The problem of structural representation of a protein until now has been tackled by specific descriptors, usually point-based and not suited for management and processing. Among these we can quote: spin image [26], [4], context shape [13], harmonic shape [14] and PGI [7]. Following [8], in this paper we propose to model the 3D structure of each protein employing its concavities and organize them in a hierarchical structure, the Protein Concavity Tree (PCT), that represents its concavity tree ([1]). For the learning task needed to classify proteins represented in such a way, we propose to employ the Graph Neural Network (GNN)[27] model that is particularly suited for learning in the structured domain. Graph classification techniques have been successfully employed in many applications fields. Borgwardt et al. [6] applied the graph kernel method to classify protein 3D structures. It outperformed classical alignment-based approaches. Karklin et al. [20] built a classifier for non-coding RNAs employing a graph representation of RNAs.

The remainder of the paper is as follows. In Section 2 the PCT data structure is presented. In Section 3 the Graph Neural Network model is introduced, while in Section 4 experimental results are presented. Section 5 concludes the paper.

2 Protein Concavity Tree

The concavity tree data structure is used for describing non-convex two and three dimensional shapes. It is a rooted tree in which the root represents the convex hull of the object and each node describes the set of objects obtained by subtracting the object from the convex hull. A leaf of the tree corresponds to a convex shape. Each node in a concavity tree stores information related to a concavity and to the tree meta-data (i.e. the level of the node, height, number of nodes, and number of leaves in the subtree rooted at the node, etc.).

In order to employ such structure to describe a protein, it is necessary to introduce some molecular models of a protein. The simplest model represents atoms as hard spheres whose radius, namely the van der Waals radius (VDWR), indicates the largest distance at which an atom repels its neighbors. The union of these hard spheres is called van der Waals volume and the resulting enclosing surface is termed the van der Waals surface (VDWS). The Solvent Accessible Surface (SAS) is the locus of the centers of a spherical probe that rolls over the molecular system. Geometrically, it coincides with the VDWS of the system where VDWR is increased by the size of the radius of the probe. The Solvent Excluded Surface (SES), often identified with the Molecular Surface, separates the volume accessible to a finite size solvent probe from the inaccessible one. This definition, based on a hard sphere model of both the solute and the solvent, was suggested by Lee and Richards [18]. An example of the surfaces, in the 2D case, is showed in Fig. 1.

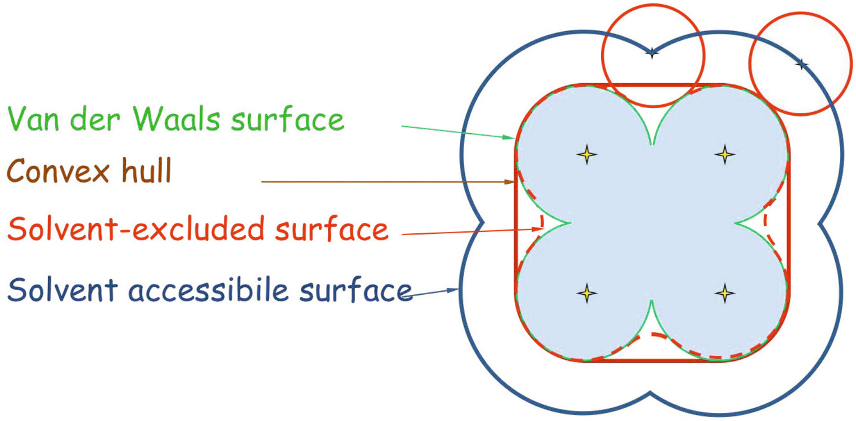

Fig. 1. Examples of molecular models of a protein in 2D

In the discrete space the protein and its convex hull (CH) are defined in a 3D grid of dimension $L \times M \times N$ voxels. The grid is extended one voxel beyond the minimum and maximum coordinate of the SES in each orthogonal direction, so that both the SES and the CH are inside the grid. Let us call R the volume of the concavity [5] between the CH and the SES

$$R = CH \cap \overline{SES} \tag{1}$$

where $\overline{SES}$ identifies the complement of the SES. Let B_{CH} the set of the border voxels of CH

$$B_{CH} = CH - [CH \bullet K] \tag{2}$$

where $\bullet$ represents the erosion operator of mathematical morphology and K the discrete unit sphere (in the discrete space a $3 \times 3 \times 3$ cube). Starting from the voxels in B_{CH}, the propagation algorithm proposed in [9] computes the connected components A of R that represent both pockets and tunnels. In order to separate the different pockets and tunnels, A is partitioned into a set of disjoint segments $P_{SES} = \{P_1, \ldots, P_j, \ldots, P_N\}$ such that

$$P_i \cap P_j = \emptyset \qquad i \neq j \tag{3}$$

$$\bigcup_{i=1}^{n} P_i = A \tag{4}$$

In order to retain the topological relationships between the partitioned segments, we introduce a novel structural description of the protein called Protein Concavity Tree. The PCT of a given protein is computed by recursively applying the segmentation process, where at each stage, exact measures of the remaining concavities can be computed. Once the complex shape is segmented into a set of pockets, each pocket can be subsequently decomposed into simpler regions. The process continues until all regions are convex. In this way, the complete description is given in terms of the region's features and their spatial relationship.

The concavities at each level of the PCT can be analyzed and described using many different features (computed in terms of the voxels belonging to the concavities). In this paper we have selected the following discriminative features: Pocket Volume [19], given by the number of voxels belonging to the pocket; Pocket Surface-to-Volume Ratio, where the pocket surface is computed as the number of SES voxels belonging to the pocket; Skewness and Kurtosis of Height Distribution, that compute the asymmetry of the surface of the pocket and its deviation from an ideal bell–shape surface, respectively; Mouth Aperture of the pocket, computed as the ratio between the perimeter and the area of the aperture; Travel Depth [10], computed as the shortest path from the CH to the surface of the pocket. For a detailed description of these features, the reader is referred to [8].

As an example of the described process, Figures 2 and 3 show the protein 1MK5 and its concavity tree, respectively.

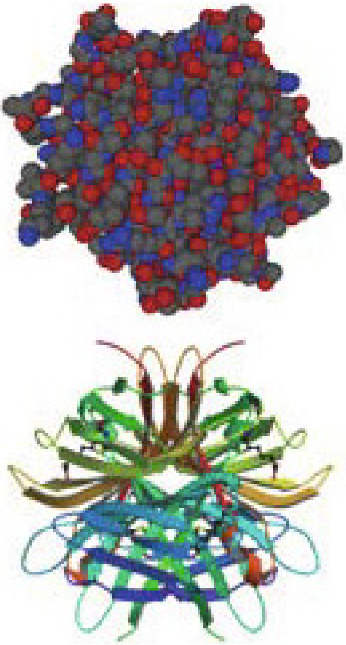

Fig. 2. Protein 1MK5

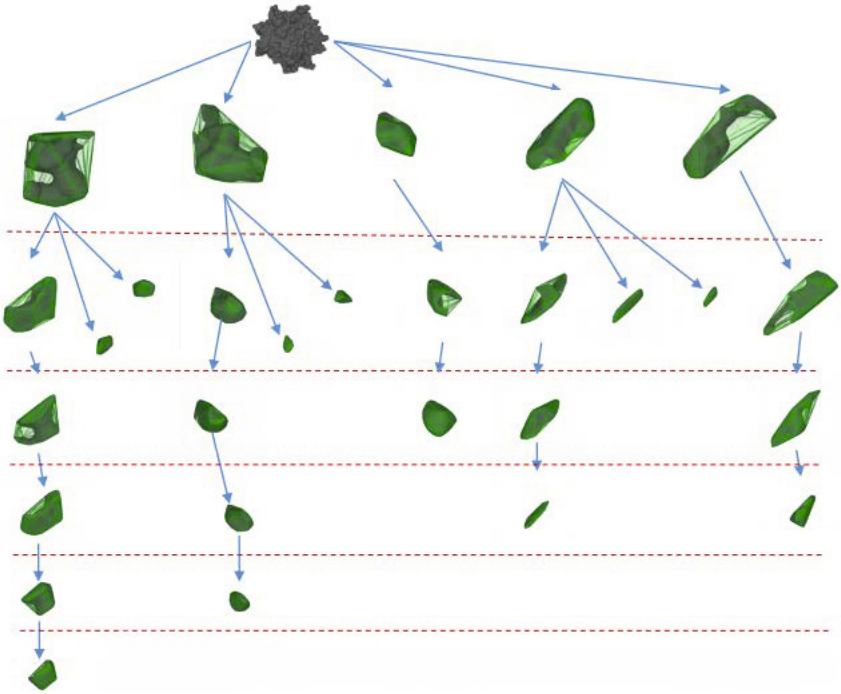

Fig. 3. Concavity Tree of protein 1MK5. In order to highlight the details a scaling factor is applied increasing the representation level.

3 Graph Neural Networks

In the recent years, many powerful machine learning methods have been developed to deal with one dimensional data, even though more complex data structures can be employed to obtain a better representation of data and possibly a better solution to a given problem. Neural methods are an example of techniques that evolved to handle structured data, where the original connectionist models have been modified to process sequences [29], trees and graphs models [15][3].

A more general supervised neural network model, is called Graph Neural Network (GNN). GNN extends Recursive Neural Network (RNN) [12] [28] since it can process a more general class of graphs including cyclic, directed, and undirected graphs, and it can deal with node-focused applications without any preprocessing steps.

The basic idea behind GNNs is the information diffusion mechanism, i.e. a graph is processed by a set of units, one for each node of the graph, connected following the graph connectivity. This representation of the graph, called encoding network, is unfolded through the structure of the input graph. At each step, all the units compute their states using information of the adjacent nodes, until stable state is reached

$$\begin{cases} x_n = f_w(l_n, l_{cp[n]}, x_{ne[n]}, l_{ne[n]}) \\ o_n = g_w(x_n, l_n) \end{cases} \tag{5}$$

where l_n is the label of node n, $l_{cp[n]}$ are the labels of its edges, $x_{ne[n]}$ are the states of the nodes in the neighborhood of n, $l_{ne[n]}$ are the labels of the nodes in the neighborhood of n, and x_n is the state of node n, i.e. it represents the concept denoted by node n. This value, along with the label l_n, is used to produce the output o_n, i.e. the decision about the concept.

The diffusion mechanism is constrained in order to ensure that a unique stable equilibrium always exists. Although this mechanism has been already used in cellular neural networks and Hopfield neural network, the one used in GNN allows the processing of more general classes of graphs. Also in [27] it is proved that GNNs show a sort of universal approximation property and they can approximate most of the practically useful functions on graphs.

4 Experimental Results

This section presents preliminary results of the proposed approach. In this paper, we use data from the spatial structures of Structural Classification of Proteins (SCOP) databases which are based on the similarities of their amino acid and three-dimensional structures [21]. This database includes fold classification family that are connected to sequence similarity. In particular, the employed dataset

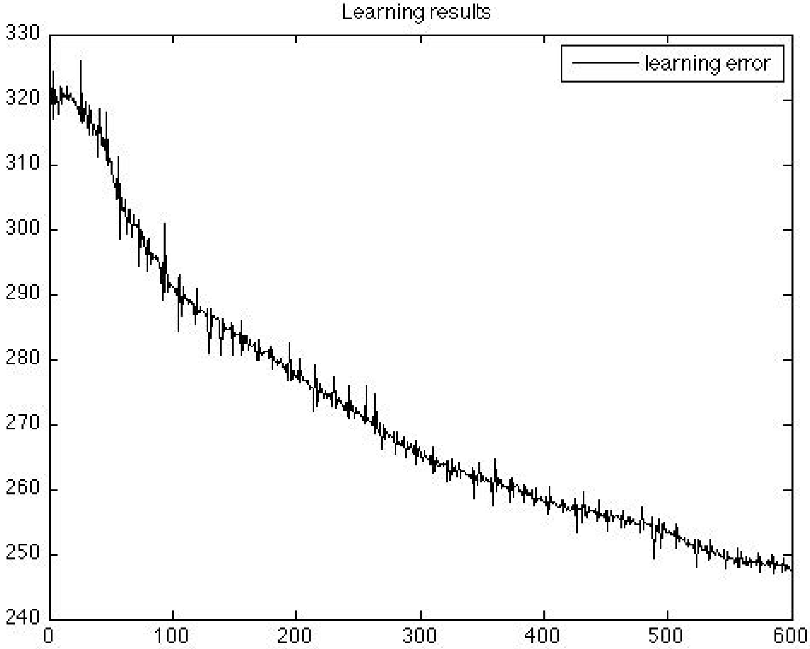

Fig. 4. GNN learning error

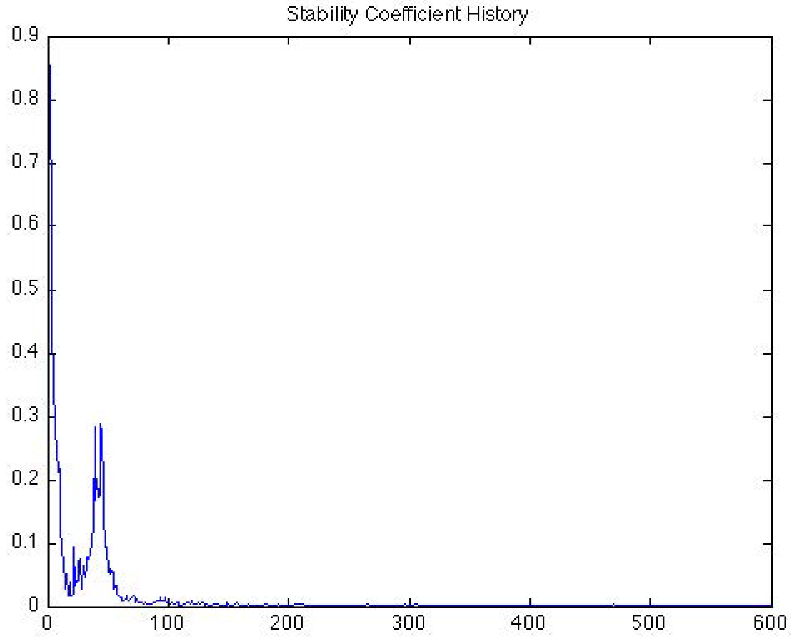

Fig. 5. GNN stability coefficient

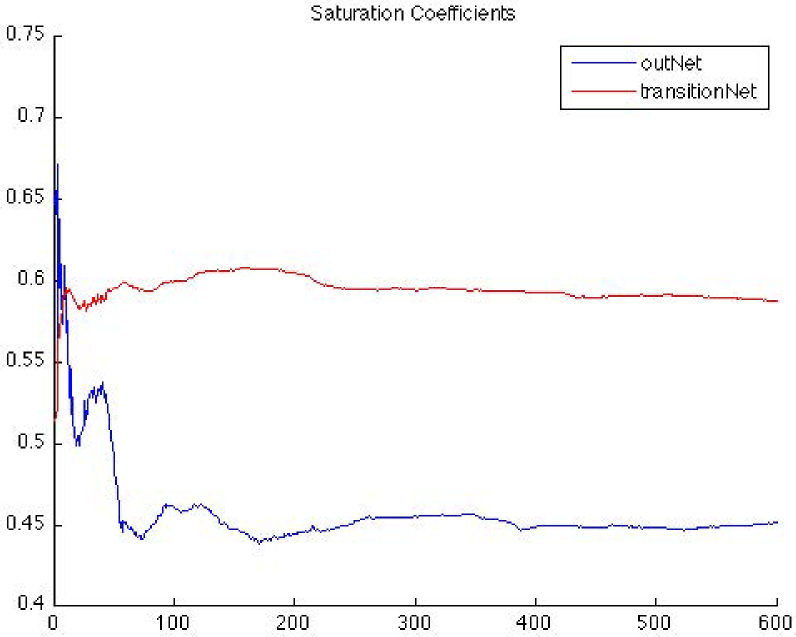

Fig. 6. GNN saturation coefficients

is composed by 70 proteins belonging to 4 classes (all alpha, all beta, alpha/beta and alpha+beta). The training set contains 90% of the samples and the testing set includes 10% of the samples, randomly selected from the whole dataset. We employed the Matlab GNN toolbox[1] using default parameters. The obtained mean accuracy is 69.93%. Although tests have not been performed on the whole SCOP database and hence comparisons with other methods have not been conducted, it is important to note how the performance of the proposed approach is very promising considering that only geometrical features have been selected. This consideration is also supported by the learning capabilities of the GNN that can be observed in the plot of the learning error, shown in Figure 4. Moreover, as can be seen in Figures 5 and 6, the low stability coefficient and the stable saturation coefficients clearly indicate that the network is able to converge.

5 Conclusions

In this paper, we have described a novel approach to classify protein structures. The proposed approach exploits the novel idea of representing a protein by means of its Protein Concavity Tree where concavities at each level are characterized by geometrical features. In order to classify proteins represented in this way, we propose to use the Graph Neural Network model that is able to exploit the topological structure induced by the concavities hierarchy. The experimental results show that our method achieves 69.93% accuracy on protein structure classification. Future works will be devoted to further analysis, more extensive experimentation and comparisons with other techniques.

References

1. Arcelli, C., Sanniti Di Baja, G.: Polygonal covering and concavity tree of binary digital pictures. In: Proceeding International Conference MECO, vol. 78, pp. 292–297 (1978)
2. Berman, H., Henrick, K., Nakamura, H.: Announcing the worldwide Protein Data Bank. Nature Structural Biology 10(12), 980–980 (2003)
3. Bianchini, M., Maggini, M., Sarti, L., Scarselli, F.: Recursive neural networks for processing graphs with labelled edges: Theory and applications. Neural Networks - Special Issue on Neural Network and Kernel Methods for Structured Domains 18, 1040–1050 (2005)
4. Bock, M.E., Garutti, C., Guerra, C.: Spin image profile: a geometric descriptor for identifying and matching protein cavities. In: Proc. of CSB, San Diego (2007)
5. Borgefors, G., Sanniti Di Baja, G.: Methods for hierarchical analysis of concavities. In: Proceedings of the Conference on Pattern Recognition (ICPR), vol. 3, pp. 171–175 (1992)
6. Borgwardt, K.M., Ong, C.S., Schönauer, S., Vishwanathan, S.V.N., Smola, A.J., Kriegel, H.-P.: Protein function prediction via graph kernels. Bioinformatics 21(suppl. 1), i47–i56 (2005)

[1] Available online at http://www.dii.unisi.it/~franco/Research/GNN.php

7. Cantoni, V., Ferone, A., Oliva, R., Petrosino, A.: Protein Gaussian Image (PGI): A protein structural representation based on the spatial attitude of secondary structure. New Tools and Methods for Pattern Recognition in Complex Biological Systems, Nuovo Cimento C 35(5, suppl. 1) (2012)
8. Cantoni, V., Gatti, R., Lombardi, L.: Proteins Pockets Analysis and Description. Bioinformatics, 211–216 (2010)
9. Cantoni, V., Gatti, R., Lombardi, L.: Analysis of geometrical and topological aptitude for protein-protein interaction. Nuovo Cimento della Società Italiana di Fisica. C, Geophysics and Space Physics 35 C, 81–88 (2012)
10. Coleman, R.G., Sharp, K.A.: Travel Depth, a New Shape Descriptor for Macromolecules: Application to Ligand Binding. Journal of Molecular Biology 362(3) (2006)
11. Day, R., Beck, D.A., Armen, R.S., Daggett, V.: A consensus view of fold space: combining SCOP, CATH, and the Dali Domain Dictionary. Protein Sci. 12(10), 2150–2160 (2003)
12. Frasconi, P., Gori, M., Sperduti, A.: A general framework for adaptive processing of data structures. IEEE Trans. on Neural Network 9(5), 768–786 (1998)
13. Frome, A., Huber, D., Kolluri, R., Bülow, T., Malik, J.: Recognizing Objects in Range Data Using Regional Point Descriptors. In: Pajdla, T., Matas, J. (eds.) ECCV 2004. LNCS, vol. 3023, pp. 224–237. Springer, Heidelberg (2004)
14. Glaser, F., Morris, R.J., Najmanovich, R.J., Laskowski, R.A., Thornton, J.M.: A Method for Localizing Ligand Binding Pockets in Protein Structures. PROTEINS: Structure, Function, and Bioinformatics 62, 479–488 (2006)
15. Gori, M., Maggini, M., Sarti, L.: A Recursive neural network model for processing directed acyclic graph with labeled edges. In: Procedings of the International Joint Conference on Neural Networks, vol. 2, pp. 1351–1355 (2003)
16. Holm, L., Kaariainen, S., Rosenstrom, P., Schenkel, A.: Searching protein structure databases with DaliLite. Bioinformatics 3, 24(23), 2780–2781 (2008)
17. Li, H., Liu, C., Burge, L., Southerland, W.: Classification of Protein 3D Structures Using Artificial Neural Network. International Journal of Machine Learning and Computing 2(6), 791–793 (2012)
18. Lee, B., Richards, F.M.: The interpretation of protein structures: Estimation of static accessibility. J. Mol. Biol. 55, 379–400 (1971)
19. Laskowski, R., Luscombe, N.M., Swindells, M.B., Thornton, J.M.: Protein clefts in molecular recognition and function. Protein Sci., 2438 (1996)
20. Karklin, Y., Meraz, R.F., Holbrook, S.R.: Classification of non-coding RNA using graph representations of secondary structure. In: Pac. Symp. Biocomput., pp. 4–15 (2005)
21. Marsolo, K., Parthasarathy, S., Ding, C.: A multi-level approach to SCOP folds recognition. In: Proc. Fifth IEEE Symposium on Bioinformatics and Bioengineering, pp. 57–64. IEEE Computer Society, Washington, DC (2005)
22. 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)
23. Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., Thornton, J.M.: CATH–a hierarchic classification of protein domain structures. Structure 5(8), 1093–1108 (1997)
24. Osada, R., Funkhouser, T., Chazelle, B., Donkin, D.: Shape Distributions. ACM Transactions on Graphics 21(4), 807–832 (2002)
25. Protein Data Bank, http://www.pdb.org/

26. Shulman-Peleg, A., Nussinov, R., Wolfson, H.: Recognition of Functional Sites in Protein Structures. J. Mol. Biol. 339, 607–633 (2004)
27. Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G.: The Graph Neural Network Model. IEEE Trans. on Neural Networks 20(1), 61–80 (2009)
28. Sperduti, A., Starita, A.: Supervised neural networks for the classification of structures. IEEE Trans. Neural Network 8(2), 429–459 (1997)
29. Werbos, P.J.: Backpropagation through time: what it does and how to do it. Proceedings of the IEEE 78(10), 1550–1560 (1990)