NEW AMINO ACID INDICES BASED ON RESIDUE NETWORK TOPOLOGY

JIAN HUANG[1, 2] hjian@kuicr.kyoto-u.ac.jp
SHUICHI KAWASHIMA[3] shuichi@hgc.jp
MINORU KANEHISA[1, 3] kanehisa@kuicr.kyoto-u.ac.jp

[1]*Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho Uji, Kyoto 611-0011, Japan*
[2] *School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China*
[3]*Human Genome Center, Institute of Medical Science, University of Tokyo, Minato-ku, Tokyo 108-8639, Japan*

Amino acid indices are useful tools in bioinformatics. With the appearance of novel theory and technology, and the rapid increase of experimental data, building new indices to cope with new or unsolved old problems is still necessary. In this study, residue networks are constructed from the PDB structures of 640 representative proteins based on the distance between Cα atoms with an 8 Å cutoff. All these networks show typical small world features. New amino acid indices, termed relative connectivity, clustering coefficient, closeness and betweenness, are derived from the corresponding topological parameters of amino acids in the residue networks. The 4 new network based indices are closely clustered together and related to hydrophobicity and β propensity. When compared with related amino acid indices, the new indices show better or comparable performance in protein surface residue prediction. Relative connectivity is the best index and can reach a useful performance with an area under the curve about 0.75. It indicates that the network property based amino acid indices can be useful complements to the existing physicochemical property based amino acid indices.

Keywords: amino acid index; residue network; connectivity; closeness; betweenness; clustering coefficient.

1. Introduction

Any given property of amino acids can be represented by a set of 20 numerical values, usually called a propensity scale or amino acid index [1-3]. As scales of different physicochemical and biochemical properties, amino acid indices have been widely used in various bioinformatics studies, such as predicting protein secondary structures [4], transmembrane sequences [5], surface [6, 7] and linear B cell epitopes [8-10]. Sometimes, however, the existing indices perform poorly [9], indicating that better methods or new amino acid indices are needed. With the appearance of novel theory and technology, and the rapid increase of experimental data, it is necessary to revise old amino acid indices, and build new ones.

Recently, graph and network theory have become a paradigm for research on complex biological systems [11-16]. Proteins have also been studied intensively as

networks formed by amino acid residues and their interactions [17-29]. To avoid confusion with protein-protein interaction networks, these networks are usually called residue networks or amino acid networks. In residue networks, nodes stand for amino acids and two nodes are linked together when the distance between the two nodes is shorter than a given threshold (see Fig. 1). Though constructed with various distance cutoffs and based on different residues or atoms, all residue networks studied so far have small world features [17-29]. Nearly all these networks have a normal degree distribution rather than a scale-free power-law degree distribution; though the latter is often seen in other biological networks [12, 15]. It is also very interesting that topological parameters of residue networks have shown relationships to protein folding [17-19], dynamics [23], stability [25], functional sites and residues [21, 22]. Therefore, topological properties of amino acids in residue networks may play an important role in exploring protein structure and function.

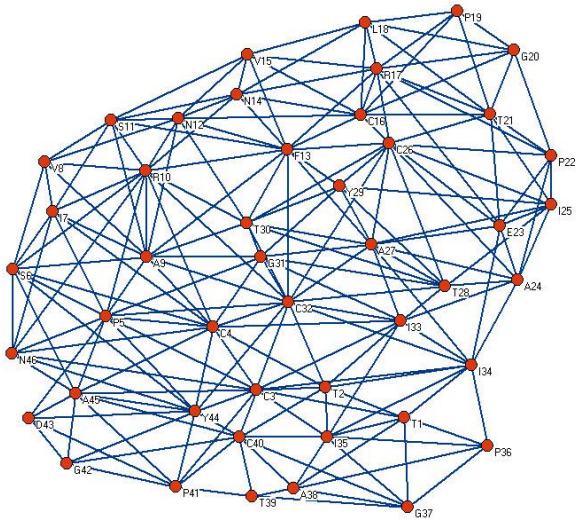

Fig 1. Residue network of crambin. Constructed from the PDB structure 1CRN, based on the distance between Cα atoms with an 8 Å cutoff and visualized with the Pajek program [30].

As reported previously, our group began to construct and maintain a database of amino acid indices almost 20 years ago [1-3]. However, most existing amino acid indices, if not all, are derived from physicochemical properties of amino acids, such as size, charge, polarity and hydrophobicity. The topological properties of amino acids in the residue networks have not been adequately studied and included into the AAindex database [3].

In this study, we built new amino acid indices based on the local and global topology of residue networks. We also studied the relation between these topological properties and the physicochemical properties of amino acids through cluster analysis with existing indices in the AAindex database [3]. The application of these new indices is demonstrated in protein surface residue prediction.

2. Methods and Data Sets

2.1. *Constructing residue networks*

A set of 640 representative proteins were selected from the PDB release #2006_08_20 through the web interface of the PDB-REPRDB database [31]. All the structures are determined with X-ray diffraction at a resolution of less than 2 Å and with a sequence longer than 40 residues. Structures with Cα or backbone coordinates only, or with more than one chain, or with any chain break, fragment, mutant or non-standard residue are excluded. To eliminate sequential or structural homology among the selected structures, their sequence identity and Root Mean Square Deviation (RMSD) are required to be below 30% and above 10 Å respectively.

Each residue in a structure is considered as a node. Two nodes are linked together when the Euclidian distance between their Cα atoms is shorter than 8 Å. The calculation of the Euclidian distance between two atoms has been described in detail elsewhere [32]. For each structure, a corresponding residue network is constructed.

2.2. Analyzing the topology of residue networks

The topology of a residue network can be characterized with local and global parameters. Connectivity, which is often termed "degree", is one of the most important local parameters. The connectivity of residue *r* (*Kr*) is the number of neighbors linked to residue *r*.

Clustering coefficient is another local parameter of a residue network. The clustering coefficient of residue *r* (*Cr*) reflects the probability that the neighbors of residue *r* are also neighbors of each other. In residue network, if residue *r* has *Kr* neighbors, the number of all possible links among these neighbors is *Kr (Kr-1)/2*. However, the actual links among these neighbors are counted and represented by "*Ar*". Then *Cr* is given as:

$$Cr = \frac{Ar}{Kr(Kr-1)/2} = \frac{2Ar}{Kr(kr-1)} \tag{1}$$

In network theory, closeness is a global measure for centrality. The closeness of residue *r* (*Or*) to other residues in the residue network is defined as:

$$Or = N - 1 \Big/ \sum_{s \in V, r \neq s} D(r,s) \tag{2}$$

where *N* is the network size. *D(r,s)* is the shortest path between residue *r* and another residue *s*. *V* is the set of all residues in the network.

Betweenness is another global centrality measure of a node within a network. Nodes that lie on many shortest paths between other nodes have higher betweenness than those that do not. The betweenness of residue *r* (*Br*) in the residue network is defined as:

$$Br = \sum_{q \neq r \neq s \in V} \frac{D(q,r,s)}{D(q,s)} \Big/ (N-1)(N-2) \tag{3}$$

where *D (q,r,s)* is number of shortest path between residue *q* and *s* pass through *r*, *D(q,s)* is all shortest paths between residue *q* and *s*. *(N-1)(N-2)* equals to the number of ordered pairs of residues not including *r*.

For each residue in all the residue networks, the four topological parameters are computed. For each residue network, the average parameters are calculated and the connectivity distribution is analyzed. The diameter and average path length of each residue network are also computed. These parameters are further analyzed together with protein size and structural class.

2.3. Deriving new amino acid indices

For each of the 20 amino acid commonly found in proteins, its topological properties (connectivity, clustering coefficient, closeness and betweenness) are averaged over all the residue networks constructed above. A set of 20 values for each topological parameter makes the raw amino acid index. All the 4 newly derived, raw amino acid indices are then normalized with

$$Rx = Ox / M\ x \tag{4}$$

where *Ox* is the original value of the raw amino acid index and *Mx* is the mean of that index set. The set of normalized results *Rx* makes the new, relative amino acid index based on a topological property.

2.4. *Clustering new indices with existing indices*

Hierarchical cluster analysis is applied to explore the relationships between these network based new amino acid indices and the 494 published indices in the AAindex database [3]. This is done with the program Amino Acid Explorer [33], which is based on a method reported previously by our group [1, 2].

A similar analysis is done on the 4 new indices and 11 highly related indices, which are clustered into the same branch. A minimum spanning tree is also built from the 4 new indices and 67 other indices that contain any of the following strings "hydroph", "polar," "size," "volume," "charge," and "electr" in their description.

2.5. *Predicting surface residues with new and related indices*

Seven amino acid indices are used to predict protein surface residues on 3 data sets. The first data set consists of 640 representative proteins, from which the new indices are derived. The second data set has 25 representative proteins, randomly picked from data set 1. The third data set has 25 representative proteins also. However, they are selected

from newly released PDB structures, fulfilling the previously described requirements for data set 1.

Surface residues are assigned based on their solvent accessible area at different cutoff values of 1, 10, 20, 50 and 100 $Å^2$. The solvent accessible area is computed with the NACCESS program [34] using default parameters.

The amino acid indices tested include the 4 new indices and three related indices. The new indices are Relative Connectivity (*Rk*), Relative Clustering Coefficient *(Rc*), Relative Closeness (*Ro*) and Relative Betweenness (*Rb*). The three related indices are 8 Å contact number (*N8*) [6], Parker's hydrophilicity (*Ph*) and Levitt's index (*Li*) [8]. *N8* was clustered close together with the 4 network based indices and had good performance in surface residue prediction. *Ph* was derived from experiment data and related to surface residues. Both *Ph* and *Li* have been confirmed to be one of the best indices for B cell epitope prediction [10]. If an index correlates negatively to the surface possibility, it is multiplied by -1 when used in predicting surface residues.

The prediction is completed with the classical sliding window method. In brief, a window slides from the N-terminal to C-terminal of the query protein sequence. The mean propensity value of the window is then assigned to the residue in the middle of the window. At the N- and C- termini, we use asymmetric windows to avoid omitting prediction examples. Different window sizes of 1, 3, 5, 7, and 9 are tested.

Receiver Operating Characteristics (ROC) curves are constructed and visualized with the ROCR package [35]. The area under the curve (*Aroc*) is used to evaluate the performance of each prediction [36].

3. Results

3.1. Residue networks are small worlds

All the residue networks constructed show typical small world features such as short average path length and high clustering coefficient. The connectivity distribution in residue network is normal rather than power-law (see Fig. 2) and the average path length scales up logarithmically with the network size.

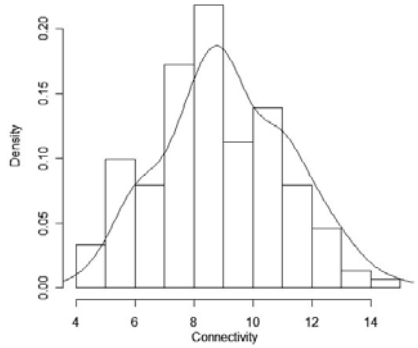

Fig 2. Connectivity distribution of sperm wale myoglobin. Residue network based on PDB structure 1A6M is shown as an example, which is in agreement with normality by Shapiro-Wilk test ($P = 0.008$).

While diameter and connectivity logarithmically scale up with residue network size, closeness, betweenness and clustering coefficient scale down logarithmically. The 640 representative proteins are assigned to "all α", "all β", "α/β", "α+β" and "others" class according to the SCOP database [37]. All the network parameters studied above show no significant differences among different structural classes.

3.2. Topologically derived new amino acid indices

Four new amino acid indices termed Relative Connectivity (*Rk*), Relative Clustering Coefficient (*Rc*), Relative Closeness (*Ro*) and Relative Betweenness (*Rb*) are derived from topological parameters of the residue networks and listed in Table 1.

Table 1. Four new amino acid indices based on residue network topology.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Rk	1.05	1.17	0.88	0.85	1.07	0.99	0.99	1.11	0.88	1.07	1.04	0.93	0.92	0.93	0.94	0.96	0.99	1.12	1.05	1.05
Rc	0.99	0.89	1.11	1.13	0.92	1.08	1.00	0.89	1.10	0.92	0.95	1.07	1.01	1.06	1.04	1.05	1.01	0.90	0.93	0.94
Ro	1.00	1.13	0.95	0.95	1.03	0.99	1.01	1.04	0.96	1.02	1.02	0.96	0.96	0.97	0.98	0.98	0.99	1.04	1.01	1.02
Rb	0.96	1.60	0.63	0.61	1.31	0.77	1.03	1.43	0.61	1.30	1.24	0.72	0.83	0.73	0.82	0.80	0.90	1.35	1.20	1.16

3.3. New indices are related to hydrophobicity and β propensity

After hierarchical clustering, all the 4 new indices are closely clustered together and related to the hydrophobicity and β propensity indices. As shown in Fig. 3, connectivity and clustering coefficient are both highly related to β propensity, and the betweenness measure is highly related to hydrophobicity. Closeness directly links to betweenness, through which the 4 new indices are joined together.

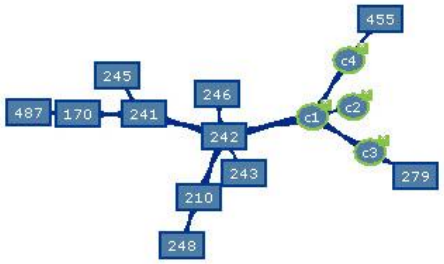

Fig 3. Minimum spanning tree built from the 4 new indices and 11 highly related indices. The new indices are displayed in circles. C1: betweenness; C2: closeness; C3: connectivity; C4: clustering coefficient. The rectangles stand for highly related indices in AAindex [3] labeled with corresponding serial number. 242: Average gain in surrounding hydrophobicity; 279: Weights for beta-sheet at the window position of 2; 455: Beta-sheet propensity derived from designed sequences.

The minimum spanning tree is also built from the 4 new indices and 67 published indices that contain any of the following string "hydroph", "polar," "size," "volume," "charge," and "electr" in their description. The close relation between the network-based

4 new indices and hydrophobicity or hydrophilicity is confirmed; in contrast, their relationships to amino acid size and charge are weak (data not shown).

3.4. Performance in predicting surface residues with new indices

Seven indices have been applied to predict surface residues on three data sets with five different sliding window sizes and surface cutoffs. Among them, relative connectivity (*Rk*) always performs best. Relative closeness (*Ro*) and relative betweenness (*Rb*) are comparable to the 8 Å contact number index (*N8*) given by Ooi *et al* [6]. Though relative clustering coefficient (*Rc*) does not perform as well as the other 3 new indices, it is still better than Parker's hydrophilicity (*Ph*) and Levitt's index (*Li*) [8] (see Fig. 4).

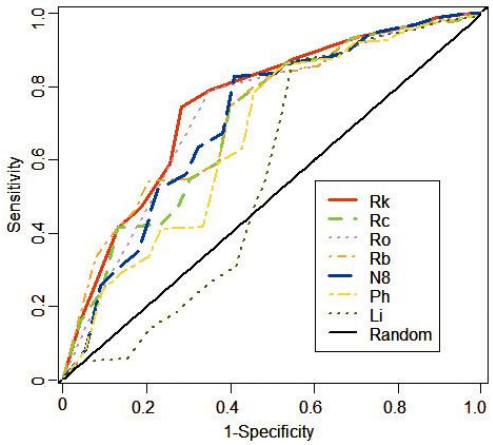

Fig 4. ROC curves for seven indices. The curves above are constructed from predictions on Data set 3 with sliding window size 1 and surface cutoff 100 Å^2. In this condition, the *Aroc* of *Rk* is about 0.75; its accuracy, sensitivity and specificity can reach 72%, 74% and 72% respectively. For a random prediction, *Aroc* is 0.5; for a perfect method, *Aroc* is 1; *Aroc* value higher than 0.7 is usually considered as a useful prediction performance.

When the size of sliding window decreases, all indices except *Li* perform better (result not shown). When the surface cutoff increases, 4 new indices tend to perform better; but the performance of *N8*, *Ph* and *Li* tend to decrease (see Table 2).

Table 2. *Aroc* from predictions on data set 3 with window size 1 and various surface cutoffs (1, 10, 20, 50 and 100 Å^2)

	SC1	**SC10**	**SC20**	**SC50**	**SC100**
Rk	0.734	0.734	0.735	0.736	0.753
Rc	0.703	0.703	0.700	0.693	0.705
Ro	0.720	0.720	0.722	0.721	0.727
Rb	0.710	0.713	0.713	0.708	0.722
N8	0.721	0.716	0.717	0.715	0.711
Ph	0.683	0.677	0.671	0.664	0.665
Li	0.658	0.635	0.628	0.600	0.557

4. Discussion

Amino acid indices are useful tools in bioinformatics. Our group has been building and maintaining a database of amino acid indices for almost 20 years [1-3]. However, most published amino acid indices, if not all, are based on physicochemical properties of amino acids, such as size, charge, polarity and hydrophobicity. Proteins can be considered as networks of amino acid residues and their interactions [17-29]. In this study, we confirmed the small world properties of such networks and built 4 new indices based on residue network topological parameters.

A very recent paper reported a very good agreement between connectivity and amino acid hydrophobicity [29]. Our results from hierarchical cluster analysis indicated that the 4 new indices do relate to hydrophobicity, but β propensity as well. As several topological parameters of residue networks have shown useful relationships to protein folding [17-19], dynamics [23], stability [25], functional sites and residues [21, 22], network topology based indices might be helpful for exploring protein structure and function.

Compared with related amino acid indices such as *Ph* and *Li*, the new indices show better performance in protein surface residue prediction. The problem of surface residue prediction is related to that of B cell epitope prediction, due to the requirement for epitopes to be surface accessible to interact with an antibody. *Ph* and *Li* have been proved to be the best two indices so far in linear B cell epitope prediction [8-10]. However, even the performance of *Ph* and *Li* are unsatisfactory [9], indicating that better methods or new amino acid indices are needed for B cell epitope prediction. Since the network topology based indices have better performance than *Ph* and *Li* in protein surface residue prediction, they might also perform better in B cell epitope prediction. This will be an area of future study for us.

In conclusion, it indicates that network topology based amino acid indices can be useful complements to the existing physicochemical property based amino acid indices.

Acknowledgments

We thank Dr Alex Gutteridge for copyediting the manuscript and giving help on R. This work was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology and the Japan Science and Technology Agency. The computational resource was provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University. The support from NSFC project (30600138) is also acknowledged.

References

[1] Nakai, K., Kidera, A., and Kanehisa, M., Cluster analysis of amino acid indices for prediction of protein structure and function, *Protein Eng.*, 2(2):93-100, 1988.

[2] Tomii, K. and Kanehisa, M., Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins, *Protein Eng.*, 9(1):27-36, 1996.
[3] Kawashima, S. and Kanehisa, M., AAindex: amino acid index database, *Nucleic Acids Res.*, 28(1):374, 2000.
[4] Kazemian, M., Moshiri, B., Nikbakht, H., and Lucas, C., A new expertness index for assessment of secondary structure prediction engines, *Comput. Biol. Chem.*, 31(1):44-47, 2007.
[5] Zhao, G. and London, E., An amino acid "transmembrane tendency" scale that approaches the theoretical limit to accuracy for prediction of transmembrane helices: relationship to biological hydrophobicity, *Protein Sci.*, 15(8):1987-2001, 2006.
[6] Nishikawa, K. and Ooi, T., Prediction of the surface-interior diagram of globular proteins by an empirical method, *Int. J. Pept. Protein Res.*, 16(1):19-32, 1980.
[7] Nishikawa, K. and Ooi, T., Radial locations of amino acid residues in a globular protein: correlation with the sequence, *J. Biochem*, 100(4):1043-1047, 1986.
[8] Pellequer, J.L., Westhof, E., and Van Regenmortel, M. H., Predicting location of continuous epitopes in proteins from their primary structures, *Methods Enzymol*., 203:176-201, 1991.
[9] Blythe, M.J. and Flower, D.R., Benchmarking B cell epitope prediction: underperformance of existing methods, *Protein Sci.*, 14(1):246-248, 2005.
[10] Larsen, J. E., Lund, O., and Nielsen, M., Improved method for predicting linear B-cell epitopes, *Immunome Res.*, 2:2, 2006.
[11] Watts, D.J. and Strogatz, S.H., Collective dynamics of 'small-world' networks, *Nature*, 393(6684):440-442, 1998.
[12] Barabasi, A.L. and Albert, R., Emergence of scaling in random networks, *Science*, 286(5439):509-512, 1999.
[13] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U., Network motifs: simple building blocks of complex networks, *Science*, 298(5594):824-827, 2002.
[14] Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N. and Barabasi, A.L., Hierarchical organization of modularity in metabolic networks, *Science*, 297(5586):1551-1555, 2002.
[15] Barabasi, A.L. and Oltvai, Z.N., Network biology: understanding the cell's functional organization, *Nat. Rev. Genet*, 5(2):101-113, 2004.
[16] Palla, G., Derenyi, I., Farkas, I., and Vicsek, T., Uncovering the overlapping community structure of complex networks in nature and society, *Nature*, 435(7043):814-818, 2005.
[17] Vendruscolo, M., Paci, E., Dobson, C. M., and Karplus, M., Three key residues form a critical contact network in a protein folding transition state, *Nature*, 409(6820):641-645, 2001.
[18] Dokholyan, N.V., Li, L., Ding, F., and Shakhnovich, E.I., Topological determinants of protein folding, *Proc. Natl. Acad. Sci. USA*, 99(13):8637-8641, 2002.

[19] Vendruscolo, M., Dokholyan, N.V., Paci, E., and Karplus, M., Small-world view of the amino acids that play a key role in protein folding, *Phys. Rev. E Stat. Nonlin. Soft. Matter Phys.*, 65(6 Pt 1):061910, 2002.
[20] Greene, L.H. and Higman, V.A., Uncovering network systems within protein structures, *J. Mol. Biol.*, 334(4):781-791, 2003.
[21] Wangikar, P.P., Tendulkar, A.V., Ramya, S., Mali, D.N., and Sarawagi, S., Functional sites in protein families uncovered via an objective and automated graph theoretic approach, *J. Mol. Biol.*, 326(3):955-978, 2003.
[22] Amitai, G., Shemesh, A., Sitbon, E., Shklar, M., Netanely, D., Venger, I., and Pietrokovski, S., Network analysis of protein structures identifies functional residues, *J. Mol. Biol.*, 344(4):1135-1146, 2004.
[23] Atilgan, A.R., Akan, P., and Baysal, C., Small-world communication of residues and significance for protein dynamics, *Biophys J.*, 86(1 Pt 1):85-91, 2004.
[24] Bagler, G. and Sinha, S., Network properties of protein structures, *Physica A*, 346(1-2):27-33, 2005.
[25] Brinda, K.V. and Vishveshwara, S., A network representation of protein structures: implications for protein stability, *Biophys J.*, 89(6):4159-4170, 2005.
[26] Kundu, S., Amino acid network within protein, *Physica A*, 346(1-2):104-109, 2005.
[27] Aftabuddin, M. and Kundu, S., Weighted and unweighted network of amino acids within protein, *Physica A*, 369(2):895-904, 2006.
[28] Aftabuddin, M. and Kundu, S., Hydrophobic, hydrophilic and charged amino acid networks within Protein, *Biophys J*, 93(1):225—231, 2007.
[29] Alves, N. A. and Martinez, A. S., Inferring topological features of proteins from amino acid residue networks, *Physica A*, 375(1):336—344, 2007.
[30] Batagelj, V. and Mrvar, A., Pajek - Analysis and Visualization of Large Networks, *Graph Drawing: 9th International Symposium*, 477, 2002.
[31] Noguchi, T. and Akiyama, Y., PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB) in 2003, *Nucleic Acids Res.*, 31(1):492-493, 2003.
[32] Huang, J., Gutteridge, A., Honda, W., and Kanehisa, M., MIMOX: a web tool for phage display based epitope mapping, *BMC Bioinformatics*, 7:451, 2006.
[33] Bulka, B., desJardins, M. and Freeland, S.J., An interactive visualization tool to explore the biophysical properties of amino acids and their contribution to substitution matrices, *BMC Bioinformatics*, 7:329, 2006.
[34] Hubbard, S. J. and Thornton, J. M., NACCESS, *Department of Biochemistry and Molecular Biology, University College London*, 1993.
[35] Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T., ROCR: visualizing classifier performance in R, *Bioinformatics*, 21(20):3940-3941, 2005.
[36] Swets, J. A., Measuring the accuracy of diagnostic systems, *Science*, 240(4857):1285-1293, 1988.
[37] Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C., SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Biol.*, 247(4):536-540, 1995.