Classification of Proteins Based on Similarity of Two-Dimensional Protein Maps

Birgit Albrecht, Guy H. Grant, Cristina Sisu, W. Graham Richards

▶ To cite this version:

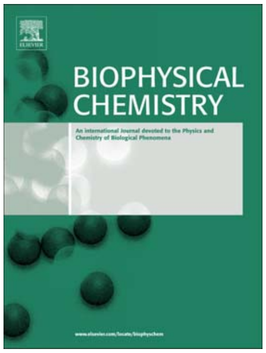
Birgit Albrecht, Guy H. Grant, Cristina Sisu, W. Graham Richards. Classification of Proteins Based on Similarity of Two-Dimensional Protein Maps. Biophysical Chemistry, 2008, 138 (1-2), pp.11. 10.1016/j.bpc.2008.08.004 . hal-00501715

HAL Id: hal-00501715
https://hal.science/hal-00501715

Submitted on 12 Jul 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Classification of Proteins Based on Similarity of Two-Dimensional Protein Maps

Birgit Albrecht, Guy H. Grant, Cristina Sisu, W. Graham Richards

PII: S0301-4622(08)00165-8
DOI: doi: 10.1016/j.bpc.2008.08.004
Reference: BIOCHE 5144

To appear in: *Biophysical Chemistry*

Received date: 3 April 2008
Revised date: 16 August 2008
Accepted date: 16 August 2008

Please cite this article as: Birgit Albrecht, Guy H. Grant, Cristina Sisu, W. Graham Richards, Classification of Proteins Based on Similarity of Two-Dimensional Protein Maps, *Biophysical Chemistry* (2008), doi: 10.1016/j.bpc.2008.08.004

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Classification of Proteins Based on Similarity of Two-Dimensional Protein Maps

Birgit Albrecht [a], Guy H. Grant [b,*], Cristina Sisu [b] and
W. Graham Richards [a]

[a] *Department of Chemistry, University of Oxford, Central Chemistry Laboratory, South Parks Road, Oxford, OX1 3QH, UK*

[b] *Unilever Centre for Molecular Informatics, University of Cambridge, Department of Chemistry, Lensfield Road, Cambridge, CB2 1EW, UK*

Abstract

Data reduction techniques are now a vital part of numerical analysis and principal component analysis is often used to identify important molecular features from a set of descriptors. We now take a different approach and apply data reduction techniques directly to protein structure. With this we can reduce the three-dimensional structural data into two-dimensions while preserving the correct relationships. With two-dimensional representations, structural comparisons between proteins are accelerated significantly. This means that protein-protein similarity comparisons are now feasible on a large scale. We show how the approach can help to predict the function of kinase structures according to the Hanks' classification based on their structural similarity to different kinase classes.

Key words: classification, function prediction, protein kinases, protein similarity, two-dimensional maps

---

---

[*] Corresponding author:
Email address: `ghg24@cam.ac.uk` (Guy H. Grant).

Preprint submitted to Biophysical Chemistry *28th August 2008*

1 Introduction

With recent advances in NMR spectroscopy and X-ray crystallography techniques, a wide range of protein structures is now available. According to the Protein Data Bank statistics [1] there are more then 40,000 known structures, a number that increases by 150 every month [2]. Even more, the number of kinase sequences in the protein database increased more than three fold in the past decade [3]. However, the biological classification of these structures is still a time consuming process and hence there is a great backlog of proteins whose structures have been identified, but are not categorized with respect to their function. Various protein classifications such as SCOP [4], CATH [2] and Pfam [5] have been developed towards a better understanding of the structural and evolutionary relationships between proteins of known structure [4]. At the foundation of all these classification techniques resides the traditional one-dimensional sequence alignment. Therefore little information can be derived with respect to function prediction of unknown proteins. DALI [6], an improved classification method, makes use of reduced dimensionality representations in order to compare various protein structures. The program is based on a sensitive measure of geometrical similarity defined as a weighted sum of similarities of intra-molecular distances [7]. To assess the similarity of two structures the program evaluates equivalent residue pairs. The assignment of equivalent residue pairs is a demanding problem and the scoring function is subjective only to geometrical criteria [6], [7]. Consequently, an automated tool to indicate an objective classification and make predictions about likely functions, could save valuable time in this process. The premise underlying this tool is that proteins with similar or even identical functions also have a high structural similarity in three-dimensions. Our group had previously reported [8] that a critical evaluation of structural similarity can be achieved using reduced dimensionality representations of protein structure. Hence, by calculating a reduced dimensionality map, namely a two-dimensional map, of the protein of unknown function and comparing it to maps of known proteins from different classes, it should be possible to predict function based on the similarity scores. High similarity scores indicate likely classifications. A full determination of the protein function can focus on those particular classes rather than having to explore the entire functionality space. Similar approaches that link structural similarity to protein classification and function, have been taken by the Thornton group. Jones and Thornton predicted functional sites of proteins, such as enzyme active sites or DNA binding sites using local similarities. With this, they acknowledged that the relationship between structure and function of a protein is not dependent on the overall fold of the structure but can also relate to small functional sites [9], and discuss different methods for the prediction of protein function from local structural similarities. A different study, also by the Thornton group [10], links structural similarity to CATH classifications. They investigated structural similarity for different sets of proteins within different EC classes. Furthermore they studied 31 superfamilies and found that structural diversity is great, sometimes even extending into the catalytic domain. Functional variation occurs only when proteins are distantly related and structurally significantly diverse. Protein Kinases are enzymes that facilitate the transfer of a phosphate group from a donor molecule to an amino acid residue of a protein. Most kinases are specific to phosphorylation of a specific type of residue, but some kinases are known to exhibit activity towards two amino acids. Protein kinases are involved in a wide variety of pathways in nature and have been linked to some of the major diseases; hence kinases have become the focus of pharmaceutical research. Also, kinases structural data is becoming available at increasing rates [3]. Classification of these new structures is a time consuming process and many new structures that may be interesting pharmaceutical targets are still waiting to be classified.

The first classification of kinases, introduced by Hanks [11], [12], divides the eukaryotic protein kinase superfamily into four major groups of basic structural and functional properties: AGC, CaMK, CMGC and PTK. The classification is based on the similarity in the sequence of the catalytic domain. The AGC group contains nine subcategories which include the cyclin-nucleotide dependent family (PKA and PKG), protein kinases C (PKC), the β-adrenergic receptor kinases (βARK), the ribosomal S6 kinase family and other similar kinases. The CaMK group can be divided up into three categories. It includes kinases that are regulated by calcium and calmodulin, the Snf1/AMPK family and other close relatives. The CMGC group has six subcategories and includes amongst others cyclin dependent kinases (CDK), the MAP kinase family and casein kinases. The PTK family groups together all tyrosine kinases, with a total of 23 subcategories. A more recent classification was proposed by Naumann and Matter, by means of target family landscape [13]. Their approach made use of three-dimensional molecular interaction field analysis of the ligand binding sites. The resulting classification lead to the identification of common binding patterns and specific interaction sites for particular kinase subfamily. The proposed method came in agreement with the original Hanks classification, proving once more that its framework is still adequate to describe all known kinases [3]. However the method used by Naumann and Matter is experimentally demanding, chemodata being a high requirement to obtain accurate results. Another classification was attempted by Cheek *et al.* [14]. They reorganized the kinases class in 12 groups based on fold and biochemical similarity. The new classification does not rule out the original Hanks' classification.

In the following we have investigated the application of reduced dimensionality maps and protein similarity to suggest potential functions for kinase structures. Protein similarity calculations are computationally very expensive tasks. For small molecules, it has been demonstrated that transferring a structure into a two-dimensional representation speeds up similarity evaluation [15]. We have applied this approach to proteins and used two-dimensional representations of protein structures to determine their similarity. Two-dimensional maps are generated using a novel algorithm and are related to the original three-dimensional structure via their respective distance matrices. Similarity between two proteins is then calculated introducing the concept of a distance dependent similarity field. With these tools it is possible to distinguish between misfolded and correct structures at the level of protein family.

2 Materials and Methods

2.1 Mapping

The generation of the two-dimensional protein maps is based on the optimisation of distance matrices of two-dimensional coordinates via a Monte Carlo-like technique so that the atom-atom distances in the 2D map are as close as possible to those in the experimental three-dimensional structure. As in the original Monte Carlo approach [16], the algorithm samples the two-dimensional space by generating random coordinates. Inter-residue distances are calculated and serve as acceptance criteria determining if the generated coordinates will be accepted or dismissed. A flowchart of the program structure is shown in Figure 1 and Figure 2 shows an example of the structural representations during the different steps.

The generation of the two-dimensional protein maps can be divided into three stages. The *first step* comprises the creation of a two-dimensional map by projecting the original three-dimensional coordinates of the protein $C\alpha$ into a randomly chosen plane containing the z-axis. Due to the iterative nature of the following stages, the choice of the projection plane does not influence the final output.

The *second step* involves the optimization of the two-dimensional map by minimizing the differences between the distance matrices of the two-dimensional $C\alpha$ positions and the initial coordinates in three-dimensional space. This is an iterative process that requires the calculation of the errors between the inter amino acid distances in the two-dimensional map and the distances between the respective amino acids in the original three-dimensional structure. This error is used as a seed for the optimisation of the distance matrices via a Monte Carlo-like technique. The seed is used with a pseudo-random number generator (PRNG) to create new coordinates for each point on the two-dimensional map. The PRNG function of the C standard library is statistically reliable and previous studies recommend its use in Monte Carlo simulations [17].

When new coordinates have been assigned to every point in the two-dimensional map, a new set of errors between the distance matrices in two- and three-dimensions are calculated. If they are all below a certain user defined threshold, the distances between any two amino acids in the two-dimensional map are as close as possible to the original distances in the three-dimensional structure and a reduced dimensionality representation of the protein was successfully obtained. It is worth mentioning that the new map does not retain any of the original properties of the three-dimensional structure except for a very similar distance matrix. If the generated matrix fails to pass the error threshold, step two is repeated. Step two is iterated a large number of times until the generated map meets the cutoff criteria, or the maximum number of steps is reached. If the later happens, no map is generated and the user must start a new calculation, this time manually setting the value of the first random seed.

The *third step* consists of normalizing the generated map to the interval [-1,1]. Finally, a file is written with the new coordinates of each amino acid.
As in the original Monte Carlo procedure [16], there is no temperature dependent selection criteria, as the generated two-dimensional maps are only a model with no physical meaning and therefore are not dependent on the Boltzmann distribution.

2.2 Similarity

In a previous paper [8] we established a new method for calculating protein similarity using reduced dimensionality maps. The two-dimensionality map retains from the original three-dimensional structure, the amino acid distance matrix information (within certain errors as described in the previous section) and quantifiable physico-chemical properties of the amino acids (polarity, hydrophobicity, side chain surface area, etc.) in form of the distance dependent similarity field [8]. However, it has the advantage of a lower number of degrees of freedom. The maps can be compared in a pointwise fashion requiring only planar transformations such as: translation, rotation and reflection.

The concept of distance dependent amino acid similarity field ($amino\_acid_{sim}$) between two amino acids p and q was previously defined [8] as the product of a similarity index dependent on the nature of the amino acids and a Gaussian representation of the distance between them:

$$amino\_acid_{sim} = sim_{pq} \cdot G_{pq} \qquad (1)$$

with

$\mathbf{G}_{pq}$**:** single Gaussian representation for distance dependence,
$\mathbf{sim}_{pq}$**:** similarity index of the pair of amino acids to be compared.

The Gaussian representation of the distance has already been shown to be an accurate representation of distance dependence of electrostatic potentials and Carbo-Index based similarities [18]. As shown by Good *et al.*, the single term Gaussian can be represented by:

$$G = -0.9437 \cdot e^{-0.0890 \cdot r^2} \tag{2}$$

with r the distance (Angstrom) between the two Cα of the amino acids to be compared [18].

The similarity field of two proteins is then the sum of the amino acid similarity fields:

$$sim = \sum_{p=1}^{l} \sum_{q=1}^{m} sim_{pq} G_{pq} \tag{3}$$

with

p: index of amino acid in protein 1,
l: number of amino acids in protein 1,
q: index of amino acid in protein 2,
m: number of amino acids in protein 2,

and G_{pq} and sim_{pq} as previously described. To calculate the overall protein similarity a normalization factor that accounts for the self-similarities of the structures needs to be applied. This factor has the same form as the similarity field, the only difference being that similarities are calculated between the protein and itself:

$$self_{sim} = \sqrt{\left(\sum_{i=1}^{l} \sum_{j=1}^{l} sim_{ij}^{p1} \cdot G_{ij}^{p1}\right) \cdot \left(\sum_{i=1}^{m} \sum_{j=1}^{m} sim_{ij}^{p2} \cdot G_{ij}^{p2}\right)} \tag{4}$$

with

$\mathbf{G}_{ij}^{p1}$**:** single Gaussian representation for distance dependence for protein 1,
$\mathbf{G}_{ij}^{p2}$**:** single Gaussian representation for distance dependence for protein 2,
$\mathbf{sim}_{ij}^{p1}$**:** similarity index of the pair of the amino acids i and j to be compared from protein 1,
$\mathbf{sim}_{ij}^{p2}$**:** similarity index of the pair of the amino acids i and j to be compared from protein 2,

and the symbols have the same meaning as in equation 3.

Hence the formula to calculate the similarity between two protein structures is the sum of all amino acid similarities divided by a normalization factor:

$$protein_{sim} = \frac{sim}{self_{sim}} = \frac{\sum_{p=1}^{l}\sum_{q=1}^{m} sim_{pq}G_{pq}}{\sqrt{\left(\sum_{i=1}^{l}\sum_{j=1}^{l} sim_{ij}^{p1} \cdot G_{ij}^{p1}\right) \cdot \left(\sum_{i=1}^{m}\sum_{j=1}^{m} sim_{ij}^{p2} \cdot G_{ij}^{p2}\right)}} \quad (5)$$

with symbols having the same meaning as previously described.

The similarity index sim_{pq} can be chosen to account for different properties of the protein, and thus place emphasis on different topological features during similarity comparisons. We have previously tested a range of different properties for this [8]. In the following we have tried a different approach by using actual three dimensional similarities of the amino acids based on the Hodgkin index calculated by the ASP module in the Tsar computational package [19] as well as other quantifiable properties like structural similarity[20] and hydrophobicity pattern [21]. ASP is a software to calculate the similarity between two molecules originally developed by Burt and Richards [19]. The similarity index representing the nature of the amino acids was calculated as the Hodgkin similarity index [22] between the two amino acid residues. The Hodgkin index was initially developed starting from the Carbo Index [23], and accounts for the sign and magnitude of a property X of amino acids. In the present case, the Hodgkin similarity index was computed based on the electrostatic potential of amino acids:

$$Hodgkin_{AB} = \frac{2\sum_{i=1}^{N} X_{iA}X_{iB}}{\sum_{i=1}^{N}(X_{iA})^2 \cdot \sum_{i=1}^{N}(X_{iB})^2} \quad (6)$$

with

$\mathbf{X}_{iA}$: intensity of property X of amino acid i of protein A,
$\mathbf{X}_{iB}$: intensity of property X of amino acid i of protein B,
N: total number of amino acids in protein A, respectively B.

To calculate the similarity of two proteins, the two-dimensional maps of the proteins are first aligned to the origin. The position of the origin of each map is described by three coordinates x, $y \in$ [-1,1] and $\phi \in$ 0-360°. In order to cover the entire sample space in a fast, statistically reliable way a new Monte Carlo algorithm is used which offers considerable increases in the speed of the alignment. At each step the similarity of the two proteins is calculated by summing the single amino acid similarities using the equation and similarity indices described above. The initial random seed is then calculated as the difference in the similarity scores between the previous step and the current step. The seed is used to generate new values for x, y and ϕ, therefore affording a new position of one map with respect to the other. The similarity calculation is repeated 1600 times (iteration cutoff set manually). At every step, the similarity score is retained only if it is larger then the previous one. At the end of the 1600 cycles a new map is created by reflection of the first one against one of the axes. The iteration is rerun for the new map. Finally, the best similarity score is returned. The score value ranges between 0 (no similarity) and 1 (identity). A more detailed description of the mapping process and map comparison can be found in Albrecht *et al.* [8].

Structure files for the selected proteins were downloaded from the RCSB Protein Data Bank [1]. They were then converted into two-dimensional maps using a Monte Carlo mapping procedure. The similarity between the proteins was calculated based on the similarity field method described in the previous section. The assays were performed using various similarity indices and descriptors of various amino acids properties. In order to assess the validity of the presented method, several tests were carried out. Hanks classification of kinases was used as standard. Four representative sets of proteins were chosen corresponding to the four kinase classes described by Hanks [11,24]. Only protein with associated three-dimensional structural information were considered. Most of the structures in the database did not contain information as to whether they represent active or inactive conformations and hence the dataset are composed of a mixture of active and inactive structures as well as a majority of unknown conformations. Owing to the resolution of the mapping code this should not have a major influence on the results, as the granularity was chosen to disregard minor structural changes as shown previously for the NMR ensemble [8]. It should however be noted that a dataset of just active or inactive structures would have been preferable and could lead to improved results and better discrimination.

Test were performed in duplicate. For the purpose of this paper, only one set of results will be shown. Model classes containing 11 proteins each, were constructed corresponding to the four major Hanks kinase classes: AGC, CAMK, CMGC and PTK. A full account of the structures used to construct the test classes is given in Appendix (Tables 5, 6, 7 and 8). For each assay two test set proteins were selected (Table 1). Each set contained one single structure from each of the different Hanks' classes. The test classes do not contain the structures included in the test sets. All the proteins in the test sets were compared to all the proteins in each test class. Proteins scoring higher then a class cut off value are most likely to belong to that group. Class similarity cut off scores were calculated as the average similarity within the same class of proteins (Table 4). In this way it is possible to predict to which class, the query protein is most likely to belong to.

In order to asses the validity of our method, average similarity scores on the global comparison, “all against all” proteins, were also calculated using common multiple protein comparison algorithms like sequence alignment [25], Mammoth [26] and TMalign [27]. ‘All against all” comparison of the proteins was performed using reduced dimensionality representations of the proteins, having as similarity index the hydrophobicity pattern, structural similarity and side chain similarity of amino acids. The amino acid similarity matrices are given in the Appendix (Tables 9, 10 and 11).

3 Results and Discussion

Similarity calculations performed on test structures from four different Hanks classes, using as scoring index, the Hodgkin index (Figure 3) structural similarity of the amino acids (Figure 4) and amino acid hydrophobicity (Figure 5), showed good discrimination between the classes. Looking at the results of test proteins depicted in Figures 3, 4 and 5, it can be observed that for each class, the test set protein corresponding rightfully to that family, presents the highest similarity score and can be correctly classified. Depending on the scoring index used, some classes present a better discrimination than others, highlighting the fact that the amino acid properties contribute with different weight towards protein similarity. In the present case, using the Hodgkin index as the scoring matrix, the test protein from the AGC class, showed a higher similarity score with respect to CAMK proteins rather than AGC proteins. In the same manner, CMGC test proteins showed a stronger similarity with respect to AGC and CAMK protein classes, when compared on the basis of structural similarity or amino acid hydrophobicity pattern. Also, from the figures it can be observed that in all cases, both sets of test proteins gave comparable results, supporting the reproducibility and viability of the described method.

Figures 6, 7 and 8 presents the results of the similarity assay for each individual protein from the test set with respect to each class. From the graphs it can be observed that the test structures, most of the time, scored highest, in the class they rightfully belong to. This indicates a correct classification. The property scoring matrix plays an essential role in tuning the classification. Comparing the proteins on the basis of the amino acid structural similarity or hydrophobicity pattern, the test proteins from AGC, CAMK score above the GCMC class threshold. This would indicate either that structural similarity and hydrophobicity index of amino acids contribute to a lower extent in the class differentiation or that the particular proteins chosen for this test present a high similarity with respect to the hydrophobic pattern and three-dimensional structural features. Further analysis must be performed in order to obtain a better understanding.

To get a more quantitative assessment of this method the average similarities for each of the four Hanks' classes and the average similarity of all proteins considered across all four classes were calculated (Table 2). The average similarity within the four different classes is increased compared to the average similarity of all proteins, and this allows for the correct classification of structures. Figure 9 illustrates the similarities of all structures used within the four Hanks' classes. Again it can be seen that in general the similarity within the Hanks' classes is higher than between different classes. High similarities between structures from different classes are still possible but usually they are exceptions.

The enrichment factor for the prediction based on similarity over a random choice was calculated (Table 3). The enrichment factor is defined as the ratio of the probability for the correct class over the random choice probability. As there are four different classes, there is a 25% chance of picking the correct class on a random choice basis. The prediction probabilities based on the similarity scores for the different classes per test structure were also computed, so that the sum of all probabilities of one test structure across all classes adds up to 100%. Not only were the correct classes always predicted with the highest similarities, but enrichment of up to 150% can be observed depending on the structure considered.

The present method was compared to other known protein comparison techniques. The results are presented as heat maps and dendograms in Figures 10 11 and 12. In the same line using amino acid structural similarity matrix, the proteins were compared using their two-dimensional maps. Results are shown in Figure 13. The heat maps and dendograms were constructed using R [28]. Based on sequence alignment, the phylogenetic tree of the 44 protein set was constructed. It can be clearly seen that the phylogenetic relationship is not always preserved between proteins of the same class.
All methods seems to correctly classify the AGC group, indicating a strong functional and structural relationship between proteins belonging to this class. The divergence observed in clustering the other proteins can be explained by lateral similarities between proteins from different classes. On average, the three-dimensional structural classification methods out-rule the sequence alignment. The reduced dimensionality similarity method classified the proteins correctly, giving results similar to the one obtained using the three-dimensional techniques, compares well with all the other three-dimensional techniques. The advantage that the described method possess is the flexibility in choosing the similarity parameters.

4 Conclusions

With an abundance of proteins whose structures have been identified, but whose functions has yet to be characterized, a method to predict or at least indicate potential classifications and functions would be a valuable tool. The presented work makes use of protein similarity to predict classifications of protein structures. Similarity was based on two-dimensional protein maps as this offers a great increase in the speed of the calculations. We have shown that in principle the classification of unknown structures via similarity of their reduced dimensionality maps is feasible. The obvious potential pitfall of the method lies within the protein datasets used as representative groups for a protein family or class. Firstly, these sets should include a sufficient number of proteins. However, in classes and families that are of most interest there are usually only a limited number of structures available. If the dataset that is available to represent a class is small, over-expressing certain features, or bias towards structures, becomes a real issue. Even if a larger number of structures are available, care must always be taken that the selected set is representative of all possible protein constructs within the set to be examined. In the present work relatively small representative sets composed of a careful selection of different structures from different kinase classes was used. These small sets are already sufficient to give a good indication of the protein function. Even though results may not always be absolute and the top scores may not always reliably classify a structure, the overall results usually provide a very good starting point for experimental classification, as indications are provided for potential classifications and the field is sufficiently reduced. In almost all cases we managed to predict protein classes correctly. Only in a very few instances were results for two classes almost identical with a slight preference for the incorrect class. More basic function predictions on a higher level of the protein classification trees are a lot easier, not only due to the availability of more representative sets, but also because similarity between different branches is usually lower than similarity within a specific classification. We have concentrated only on kinases, which are already a small subset of all possible proteins, however even within this group we still find a good distinction between structures from different classes. The technique presented in this paper comes as an improvement of the existing classification procedures. The method is fast and simple to implement. The complicated proteins are reduced to two-dimensional map. The resulting maps bare no physical connotation in the way that they do not retain the $C\alpha$ backbone. However, each point on the map is characterized by the physico-chemical properties of its amino acid analogue. Comparison of the protein is reduced to a map comparison. Also, depending on the similarity parameters used, the comparison can account for various properties like shape, hydrophobic/hydrophilic nature, polarity, conformation, folding angles preference. With the small time requirements for this method, it can aid to experimental classification of proteins as highly likely classes can be identified very quickly. The main advantage of using the presented method would be that it can account for different physico-chemical properties of amino acids as well as three-dimensional structural information preserved in the distance matrix. It therefore offers a more reliable and unbiased differentiation between proteins. Still there remained few question unanswered, like what is the sensitivity of the method at comparing enantiomer like structures. Pilot tests performed on D– and L–monellins showed that two distinguishable maps can be obtained. A parameter that would translate the protein orientation to its bidimensional representation would help improve the discrimination. Work on this subject is in progress and results will be announced in a future paper.

References

[1] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissing, I. Shindyalov, P. Bourne, The Protein Data Bank, Nucleic Acids Research 28 (2000) 235 – 242.

[2] C. Orengo, A. Michie, S. Jones, D. Jones, M. Swindells, J. Thornton, CATH - a hierarchic classification of protein domain structures, Structure 5 (1997) 1093 – 1108.

[3] S. Cheek, K. Ginalski, H. Zhang, N. Grishin, A comprehensive update of the sequence and structure classification of kinases, BMC Structural Biology 5 (2005) 6.

[4] A. Murzin, S. Brenner, T. Hubbard, C. Chothia, SCOP: A structural classification of proteins database for the investigation of sequence and structures, Journal of Molecular Biology 247 (1995) 536 – 540.

[5] R. Finn, J. Mistry, B. Schuster-Bockler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S. Eddy, E. Sonnhammer, A. Bateman, Pfam: clans, web tools and services, Nucleic Acids Research 34 (2006) D247 – D251.

[6] L. Holm, C. Sander, Protein structure comparison by alignment of distance matrices, Journal of Molecular Biology 233 (1993) 123 – 138.

[7] L. Holm, J. Park, DaliLite workbench for protein structure comparison, Bioinformatics 16 (2000) 566 – 567.

[8] B. Albrecht, G. Grant, W. Richards, Evaluation of structural similarity based on reduced dimensionality representations of protein structure, Protein Engineering, Design and Selection 17 (2004) 425 – 432.

[9] S. Jones, J. Thornton, Searching for functional sites in protein structures, Current Opinions in Chemical Biology 8 (2004) 3 – 7.

[10] A. Todd, C. Orengo, J. Thornton, Evolution of function in protein superfamilies from a structural persepctive, Journal of Molecular Biology 307 (2001) 1113 – 1143.

[11] S. Hanks, A. Quinn, Protein kinase catalytic domain sequence database: Identification of conserved features of primary structure and classification of family members, Methods in Enzymology 200 (1991) 38 – 62.

[12] S. Hanks, T. Hunter, The Eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification, FASEB 9 (1995) 576 – 596.

[13] T. Naumann, H. Matter, Structural classification of protein kinases using 3d molecular interaction field analysis of their ligand binding sites: target family landscapes, Journal of Medicinal Chemistry 45 (2002) 2366 – 2378.

[14] S. Cheek, H. Zhang, N. Grishin, Sequence and structure classification of kinases, Journal of Molecular Biology 320 (2002) 855 – 881.

[15] B. Allen, G. Grant, W. Richards, Similarity calculations using two-dimensional molecular representations, Journal of Chemical Information and Computer Sciences 41 (2001) 330 – 337.

[16] N. Metropolis, S. Ulam, The Monte Carlo method, Journal of the American Statistical Association 44 (1949) 335 – 341.

[17] R. Cassia-Moura, C. Sousa, A. Ramos, L. Coelho, M. Valenca, Yet another application of the Monte Carlo method for modeling in the field of biomedicine, Computer Methods and Programs in Biomedicine 78 (2005) 223 – 235.

[18] A. Good, W. Richards, Rapid evaluation of shape similarity using Gaussian functions, Journal of Chemical Information and Computer Sciences 33 (1993) 112 – 116.

[19] C. Burt, W. Richards, The application of molecular similarity calculations, Journal of Computational Chemistry 11 (1990) 1139 – 1146.

[20] K. Niefind, D. Schomburg, Amino acid similarity coefficients for protein modeling and sequence alignment derived from main-chain folding angles, Journal of Molecular Biology 219 (1991) 481 – 497.

[21] P. Riek, M. Handschumacher, S.-S. Sung, M. Tan, M. Glynias, M. Schluchter, J. Novotny, R. Graham, Evolutionary conservation of both the hydrophilic and hydrophobic nature of transmembrane residues, Journal of Theoretical Biology 172 (1995) 245 – 258.

[22] E. Hodgkin, W. Richards, Molecular similarity base on electrostatic potential and electric field, International Journal of Quantum Chemistry. Quantum Biology Symposium 14 (1987) 105 – 110.

[23] R. Carbo, L. Leyda, M. Arnau, How similar is a molecule to another? An electron density measure of similarity between two molecular structures, International Journal of Quantum Chemistry 17 (1980) 1185 – 1189.

[24] C. Smith, I. Shindyalov, S. Veretnik, M. Gribskov, S. Taylor, L. ten Eyck, P. Bourne, The protein kinase resource, Trends in Biochemical Sciences 22 (1997) 444 – 446.
URL `http://www.kinasenet.org/pkr`

[25] M. Larkin, G. Blackshields, N. Brown, R. Chenna, P. McGettigan, H. McWillam, F. Valentin, I. Wallace, A. Wilm, R. Lopez, J. Thompson, T. Gibson, D. Higgins, Clustal W and Clustal X version 2.0, Bioinformatics 23 (2007) 2947 – 2948.

[26] A. R. Ortiz, C. Strauss, O. Olmea, MAMMOTH (Matching molecular models obtained from theory): An automated method for model comparison, Protein Science 11 (2002) 2606 – 2621.

[27] Y. Zhang, J. Skolnick, TM-align: a protein structure alignment algorithm based on the TM-score, Nucleic Acids Research 33 (2005) 2302 – 2309.

[28] K. Hornik, The R FAQ (2008).
URL `http://CRAN.R-project.org/doc/FAQ/R-FAQ.html`

5 Tables and Figures

Figure 1. Flowchart of algorithm for the two-dimensional mapping of protein structures. An initial projection generates starting coordinates, which are subsequently optimized using a Monte Carlo algorithm.

Figure 2. Snapshots during the generation of a Monte Carlo two-dimensional map. The three-dimensional coordinates are extracted from the original three-dimensional structures and use to generate initial two-dimensional starting coordinates via a projection. These two-dimensional coordinates are then optimised using a Monte Carlo algorithm so that the two-dimensional distance matrix converges to the initial three-dimensional matrix to give the final two-dimensional map.

Figure 3. Average similarity score of test structures of each class with all four different classes using Hodgkin similarity index. The different colours account for the different test set (green - test set 1, blue - test set 2).

Figure 4. Average similarity score of test structures of each class with all four different classes using structural similarity of the amino acid as the similarity index. The different colours account for the different test set (yellow - test set 1, red - test set 2).

Figure 5. Average similarity score of test structures of each class with all four different classes using the amino acid hydrophobicity pattern as the similarity index. The different colours account for the different test set (red - test set 1, blue - test set 2).

Figure 6. Scores of proteins form test set 1 in the four different classes. Similarities were calculated using Hodgkin similarity of amino acid residues as similarity index. The average similarity of each test structure is represented by a column. The structures used for the datasets are listed in Tables 5 - 8.

Figure 7. Scores of proteins from test set 1 in the four different classes. Similarities were calculated using structural similarity of the amino acid residues as the similarity index. The average similarity of each test structure is represented by a column. The average similarity within each class is indicated by a line. The structures used for the datasets are listed in Tables 5 - 8.

Figure 8. Scores of proteins from test set 1 in the four different classes. Similarities were calculated using hydrophobicity pattern of the amino acid residues as the similarity index. The average similarity of each test structure is represented by a column. The average similarity within each class is indicated by a line. The structures used for the datasets are listed in Tables 5 - 8.

Figure 9. Similarities matrix for all structures across the four different Hanks classes. The numbers in the boxes indicate the average similarity between structures from the two relevant classes. In general higher similarities are observed within one class, but occasional high similarities can occur between structures from different Hanks classes.

Figure 10. Heat map and dendrogram classification of test set protein kinases using Clustal-X [25]. Colour code: zero similarity - black, identity - white.

Figure 11. Heat map and dendrogram classification of test set protein kinases using Mammoth-multiple [26]. Colour code: zero similarity - black, identity - white.

Figure 12. Heat map and dendrogram classification of test set protein kinases using TM-align [27]. Colour code: zero similarity - black, identity - white.

Figure 13. Heat map and dendrogram classification of test set protein kinases using structural similarity of amino acids as scoring matrix. Colour code: zero similarity - black, identity - white.

Class	Test set1	Test set2
AGC	1ATP	1JLU
CAMK	1P38	1A9U
CMGC	1E1V	1B38
PTK	1IR3	1GAG

Table 1
Table of selected test structures for the different kinase classes. Both sets of these test structures were compared to their own and the all other different classes.

Class	Average similarity in class
AGC	34%
CAMK	38%
CMGC	45%
PTK	43%

Table 2
Table of average similarities in the different Hanks' classes and of all proteins considered across the different classes. The increase in average similarities within the different classes as compared to the average across all classes is significant enough to allow for the classification of structures.

Structure	Random choice probability	Prediction probability	Enrichment factor
AGC1	25%	30%	119%
CAMK1	25%	37%	147%
CMGC1	25%	38%	151%
PTK1	25%	34%	135%
AGC2	25%	30%	119%
CAMK2	25%	36%	146%
CMGC2	25%	38%	152%
PTK2	25%	34%	134%

Table 3
Table of probabilities for the correct classification of test structures. As there are four different classes the random choice offers a 1 in 4 chance of picking the right class. The prediction based on similarity greatly improves this probability.

Kinase	Cut off scores		
	Hodgking index	Structural similarity of amino acids	Hydrophobicity pattern
AGC	0.387	0.949	0.975
CAMK	0.482	0.937	0.956
CMGC	0.438	0.820	0.878
PTK	0.425	0.852	0.926

Table 4
Average class similarity cut off scores.

PDB ID	Resolution	active/inactive
1APM	2.00	unknown
1ATP	2.20	active
1BKX	2.60	unknown
1CDK	2.00	unknown
1CTP	2.90	unknown
1FMO	2.20	unknown
1BX6	2.10	unknown
1JBP	2.20	unknown
1JLU	2.25	unknown
1O6K	1.70	unknown
1O6L	1.60	unknown

Table 5
Structures included in the AGC dataset.

PDB ID	Resolution	active/inactive
1A06	2.50	inactive
1A9U	2.50	unknown
1BL6	2.50	unknown
1BL7	2.50	unknown
1BMK	2.40	unknown
1DI9	2.60	unknown
1IA8	1.70	unknown
1IAN	2.00	unknown
1JNK	2.30	inactive
1KV1	2.50	unknown
1P38	2.10	unknown

Table 6
Structures included in the CAMK dataset.

PDB ID	Resolution	active/inactive
1AQ1	2.00	unknown
1B17	1.70	unknown
1B38	2.00	unknown
1B39	2.10	unknown
1E1V	1.95	unknown
1E1X	1.85	unknown
1ERK	2.30	unknown
1FVT	2.20	unknown
1GNG	2.60	unknown
1JAM	2.18	unknown
1LR4	2.00	unknown

Table 7
Structures included in the CMGC dataset.

PDB ID	Resolution	active/inactive
1FGI	2.50	unknown
1GAG	2.70	unknown
1GJO	2.40	unknown
1HOW	2.10	active
1I44	2.40	unknown
1IR3	1.90	inactive
1IRK	2.10	unknown
1KSW	2.80	unknown
1M14	2.60	active
1M17	2.60	unknown
2SRC	1.50	unknown

Table 8
Structures included in the PTK dataset.

1.00	0.80	0.80	0.50	0.50	0.10	0.60	0.50	0.50	0.10	0.10	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.80	0.50
0.80	1.00	0.80	0.50	0.50	0.50	0.60	0.50	0.50	0.50	0.50	0.50	0.80	0.50	0.50	0.50	0.50	0.50	0.80	0.50
0.80	0.80	1.00	0.50	0.50	0.10	0.60	0.50	0.50	0.10	0.10	0.50	0.80	0.50	0.50	0.50	0.50	0.50	0.80	0.50
0.50	0.50	0.50	1.00	0.60	0.50	0.50	0.50	0.60	0.60	0.50	0.50	0.50	0.10	0.10	0.60	0.10	0.10	0.50	0.80
0.50	0.50	0.50	0.60	1.00	0.50	0.50	0.50	0.50	0.50	0.50	0.60	0.50	0.30	0.30	0.50	0.30	0.30	0.50	0.60
0.10	0.50	0.10	0.50	0.50	1.00	0.50	0.50	0.50	0.80	0.80	0.50	0.10	0.50	0.50	0.50	0.50	0.50	0.10	0.50
0.60	0.60	0.60	0.50	0.50	0.50	1.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.60	0.50
0.50	0.50	0.50	0.50	0.50	0.50	0.50	1.00	0.50	0.50	0.50	0.50	0.50	0.30	0.30	0.80	0.30	0.30	0.50	0.50
0.50	0.50	0.50	0.60	0.50	0.50	0.50	0.50	1.00	0.50	0.50	0.60	0.50	0.30	0.30	0.80	0.30	0.30	0.50	0.60
0.10	0.50	0.10	0.60	0.50	0.80	0.50	0.50	0.50	1.00	0.80	0.50	0.10	0.50	0.50	0.80	0.50	0.50	0.10	0.50
0.10	0.50	0.10	0.50	0.50	0.80	0.50	0.50	0.50	0.80	1.00	0.50	0.10	0.50	0.50	0.50	0.10	0.50	0.80	0.50
0.50	0.50	0.50	0.50	0.60	0.50	0.50	0.50	0.60	0.50	0.50	1.00	0.50	0.10	0.10	0.60	0.10	0.10	0.50	0.80
0.50	0.80	0.80	0.50	0.50	0.10	0.50	0.50	0.50	0.10	0.10	0.50	1.00	0.50	0.50	0.50	0.50	0.50	0.80	0.50
0.50	0.50	0.50	0.10	0.30	0.50	0.50	0.30	0.30	0.50	0.50	0.10	0.50	1.00	0.30	0.50	0.50	0.80	0.50	0.10
0.50	0.50	0.50	0.10	0.30	0.50	0.50	0.30	0.30	0.50	0.50	0.10	0.50	0.30	1.00	0.30	0.80	0.80	0.50	0.10
0.50	0.50	0.50	0.60	0.50	0.50	0.50	0.80	0.80	0.80	0.50	0.60	0.50	0.50	0.30	1.00	0.30	0.30	0.50	0.60
0.50	0.50	0.50	0.10	0.30	0.50	0.50	0.30	0.30	0.50	0.10	0.10	0.50	0.50	0.80	0.30	1.00	0.80	0.50	0.10
0.50	0.50	0.50	0.10	0.30	0.50	0.50	0.30	0.30	0.50	0.50	0.10	0.50	0.80	0.80	0.30	0.80	1.00	0.50	0.10
0.80	0.80	0.80	0.50	0.50	0.10	0.60	0.50	0.50	0.10	0.80	0.50	0.80	0.50	0.50	0.50	0.50	0.50	1.00	0.50
0.50	0.50	0.50	0.80	0.60	0.50	0.50	0.50	0.60	0.50	0.50	0.80	0.50	0.10	0.10	0.60	0.10	0.10	0.50	1.00

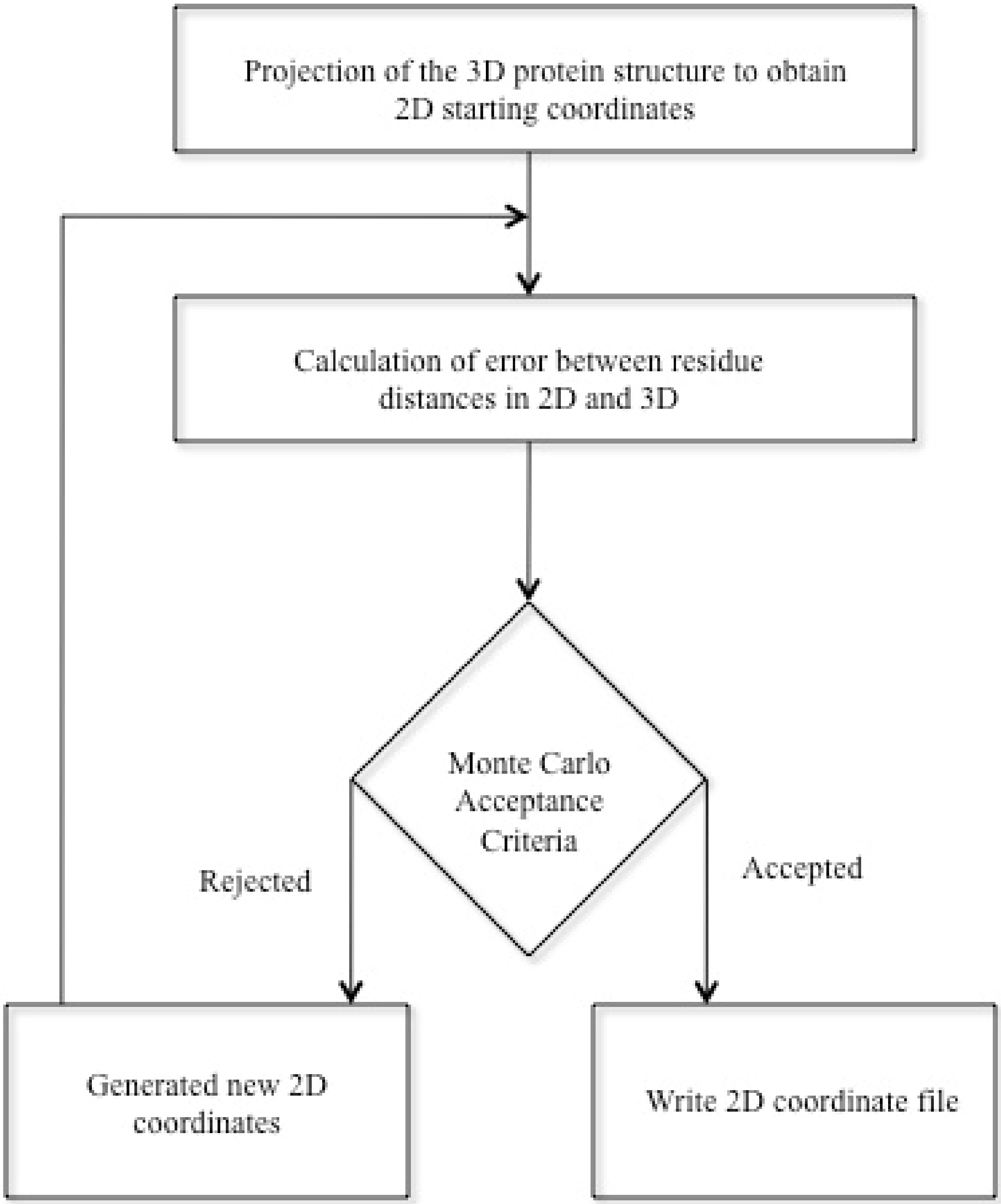
Table 9 Amino acid Hodgkin index matrix. Amino acid order: ARNDCQEGHILKMFPSTWYV.

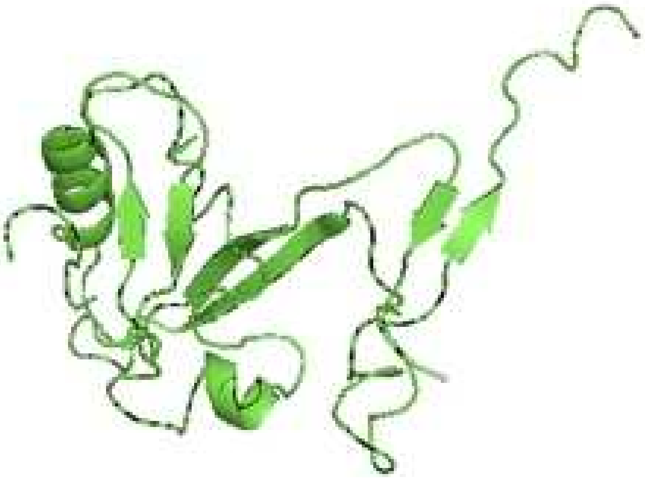
1.00	0.11	-0.31	0.07	-0.38	0.36	0.74	-0.41	-0.20	-0.26	0.54	0.60	0.43	-0.19	0.11	-0.40	-0.63	-0.11	-0.63	-0.25
0.11	1.00	-0.12	-0.08	-0.20	0.10	0.11	-0.21	0.19	0.06	0.03	0.21	-0.09	0.20	-0.21	-0.13	-0.12	0.03	-0.01	-0.04
-0.31	-0.12	1.00	0.35	0.04	-0.16	-0.27	0.40	0.04	-0.07	-0.24	-0.22	-0.13	-0.12	-0.10	0.06	0.01	-0.24	0.05	-0.16
0.07	-0.08	0.35	1.00	-0.10	0.13	0.09	0.08	0.00	-0.29	0.01	0.14	-0.08	-0.35	0.01	0.03	-0.19	-0.22	-0.22	-0.25
-0.38	-0.20	0.04	-0.10	1.00	-0.16	-0.36	0.13	-0.01	-0.16	-0.29	-0.31	-0.29	-0.07	0.14	0.35	0.15	-0.02	0.29	-0.10
0.36	0.10	-0.16	0.13	-0.16	1.00	0.42	-0.32	0.12	-0.15	0.35	0.23	0.22	-0.26	-0.18	-0.22	-0.22	-0.18	-0.32	-0.04
0.74	0.11	-0.27	0.09	-0.36	0.42	1.00	-0.46	-0.14	-0.13	0.62	0.58	0.38	-0.27	0.07	-0.43	-0.54	-0.12	-0.62	-0.09
-0.41	-0.21	0.40	0.08	0.13	-0.32	-0.46	1.00	0.09	-0.24	-0.56	-0.38	-0.30	-0.05	0.09	0.28	0.11	-0.11	0.20	-0.25
-0.20	0.19	0.04	0.00	-0.01	0.12	-0.14	0.09	1.00	0.00	-0.13	-0.26	-0.06	0.20	-0.34	-0.02	0.09	-0.02	0.14	-0.13
-0.26	0.06	-0.07	-0.29	-0.16	-0.15	-0.13	-0.24	0.00	1.00	0.27	-0.18	0.18	0.32	-0.55	-0.37	0.32	0.04	0.15	0.72
0.54	0.03	-0.24	0.01	-0.29	0.35	0.62	-0.56	-0.13	0.27	1.00	0.36	0.47	-0.08	-0.26	-0.66	-0.36	-0.15	-0.48	0.25
0.60	0.21	-0.22	0.14	-0.31	0.23	0.58	-0.38	-0.26	-0.18	0.36	1.00	0.09	-0.18	0.04	-0.21	-0.38	0.02	-0.48	-0.16
0.43	-0.09	-0.13	-0.08	-0.29	0.22	0.38	-0.30	-0.06	0.18	0.47	0.09	1.00	-0.01	-0.32	-0.53	-0.35	-0.19	-0.33	0.11
-0.19	0.20	-0.12	-0.35	-0.07	-0.26	-0.27	-0.05	0.20	0.32	-0.08	-0.18	-0.01	1.00	-0.31	-0.13	0.21	0.18	0.29	0.16
0.11	-0.21	-0.10	0.01	0.14	-0.18	0.07	0.09	-0.34	-0.55	-0.26	0.04	-0.32	-0.31	1.00	0.35	-0.21	-0.02	-0.15	-0.44
-0.40	-0.13	0.06	0.03	0.35	-0.22	-0.43	0.28	-0.02	-0.37	-0.66	-0.21	-0.53	-0.13	0.35	1.00	0.36	0.07	0.36	-0.27
-0.63	-0.12	0.01	-0.19	0.15	-0.22	-0.54	0.11	0.09	0.32	-0.36	-0.38	-0.35	0.21	-0.21	0.36	1.00	0.12	0.52	0.38
-0.11	0.03	-0.24	-0.22	-0.02	-0.18	-0.12	-0.11	-0.02	0.04	-0.15	0.02	-0.19	0.18	-0.02	0.07	0.12	1.00	0.11	-0.02
-0.63	-0.01	0.05	-0.22	0.29	-0.32	-0.62	0.20	0.14	0.15	-0.48	-0.48	-0.33	0.29	-0.15	0.36	0.52	0.11	1.00	0.17
-0.25	-0.04	-0.16	-0.25	-0.10	-0.04	-0.09	-0.25	-0.13	0.72	0.25	-0.16	0.11	0.16	-0.44	-0.27	0.38	-0.02	0.17	1.00

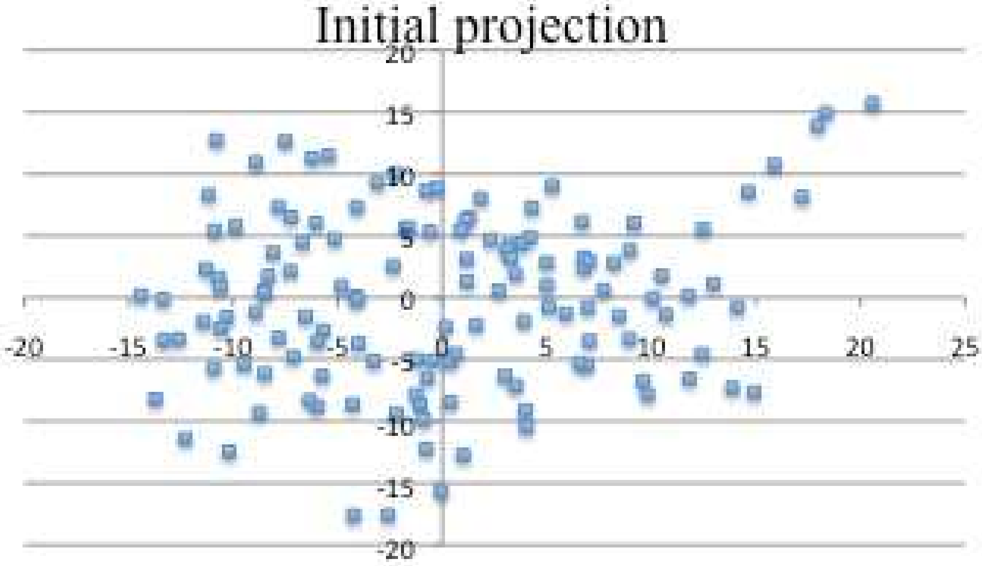
Table 10 Amino acid structural similarity index matrix. Amino acid order: ARNDCQEGHILKMFPSTWYV [20].

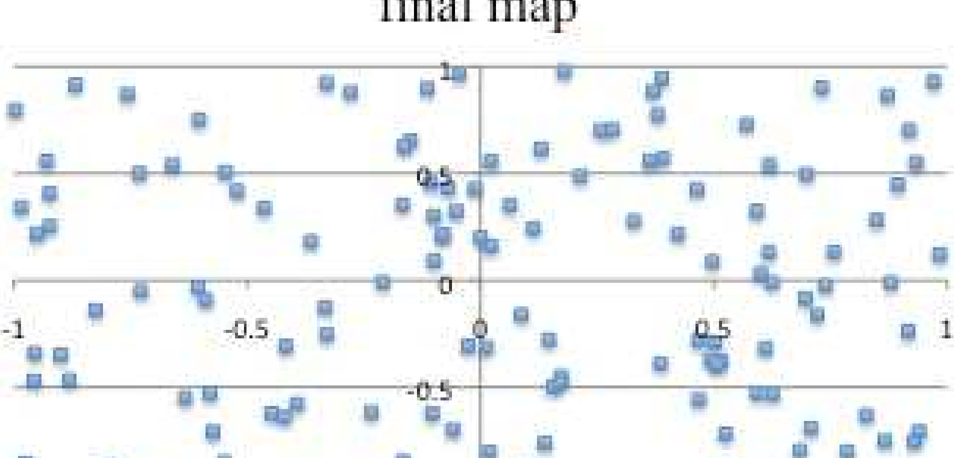
1.00	-0.07	-0.14	-0.08	0.04	-0.01	0.08	0.11	-0.11	-0.10	0.20	-0.07	0.15	-0.25	-0.07	0.14	0.11	-0.25	-0.36	0.06
-0.07	1.00	0.12	-0.01	-0.20	0.57	0.34	-0.11	0.13	-0.29	-0.20	0.47	-0.17	-0.31	-0.17	-0.16	-0.32	-0.03	-0.03	-0.29
-0.14	0.12	1.00	0.56	-0.14	0.20	0.13	0.22	0.32	-0.37	-0.39	0.22	-0.11	-0.33	0.05	0.14	0.13	-0.39	-0.27	-0.46
-0.08	-0.01	0.56	1.00	-0.24	0.19	0.37	0.11	0.22	-0.54	-0.55	0.18	-0.25	-0.40	0.09	0.22	0.23	-0.41	-0.34	-0.48
0.04	-0.20	-0.14	-0.24	1.00	-0.29	-0.51	0.25	0.18	0.26	0.23	-0.43	0.10	0.30	0.02	0.10	0.14	0.18	0.21	0.33
-0.01	0.57	0.20	0.19	-0.29	1.00	0.51	-0.13	0.05	-0.43	-0.30	0.52	-0.18	-0.44	-0.03	-0.06	-0.15	-0.17	-0.28	-0.38
0.08	0.34	0.13	0.37	-0.51	0.51	1.00	-0.23	-0.19	-0.51	-0.45	0.54	-0.14	-0.50	-0.11	-0.17	-0.27	-0.22	-0.32	-0.44
0.11	-0.11	0.22	0.11	0.25	-0.13	-0.23	1.00	0.18	-0.22	-0.14	-0.23	-0.01	-0.09	0.25	0.29	-0.02	-0.08	-0.24	-0.12
-0.11	0.13	0.32	0.22	0.18	0.05	-0.19	0.18	1.00	-0.14	-0.12	-0.09	-0.19	-0.06	0.05	0.14	0.08	-0.11	0.08	-0.18
-0.10	-0.29	-0.37	-0.54	0.26	-0.43	-0.51	-0.22	-0.14	1.00	0.60	-0.40	0.36	0.60	-0.19	-0.27	0.00	0.36	0.43	0.70
0.20	-0.20	-0.39	-0.55	0.23	-0.30	-0.45	-0.14	-0.12	0.60	1.00	-0.34	0.38	0.44	-0.20	-0.19	-0.04	0.22	0.17	0.49
-0.07	0.47	0.22	0.18	-0.43	0.52	0.54	-0.23	-0.09	-0.40	-0.34	1.00	-0.16	-0.41	-0.17	-0.24	-0.24	-0.22	-0.20	-0.36
0.15	-0.17	-0.11	-0.25	0.10	-0.18	-0.14	-0.01	-0.19	0.36	0.38	-0.16	1.00	0.22	-0.17	-0.17	-0.02	0.16	0.07	0.23
-0.25	-0.31	-0.33	-0.40	0.30	-0.44	-0.50	-0.09	-0.06	0.60	0.44	-0.41	0.22	1.00	-0.12	-0.01	0.02	0.41	0.54	0.47
-0.07	-0.17	0.05	0.09	0.02	-0.03	-0.11	0.25	0.05	-0.19	-0.20	-0.17	-0.17	-0.12	1.00	0.27	0.20	-0.13	-0.23	-0.17
0.14	-0.16	0.14	0.22	0.10	-0.06	-0.17	0.29	0.14	-0.27	-0.19	-0.24	-0.17	-0.01	0.27	1.00	0.45	-0.17	-0.21	-0.18
0.11	-0.32	0.13	0.23	0.14	-0.15	-0.27	-0.02	0.08	0.00	-0.04	-0.24	-0.02	0.02	0.20	0.45	1.00	-0.12	-0.14	0.07
-0.25	-0.03	-0.39	-0.41	0.18	-0.17	-0.22	-0.08	-0.11	0.36	0.22	-0.22	0.16	0.41	-0.13	-0.17	-0.12	1.00	0.43	0.29
-0.36	-0.03	-0.27	-0.34	0.21	-0.28	-0.32	-0.24	0.08	0.43	0.17	-0.20	0.07	0.54	-0.23	-0.21	-0.14	0.43	1.00	0.31
0.06	-0.29	-0.46	-0.48	0.33	-0.38	-0.44	-0.12	-0.18	0.70	0.49	-0.36	0.23	0.47	-0.17	-0.18	0.07	0.29	0.31	1.00

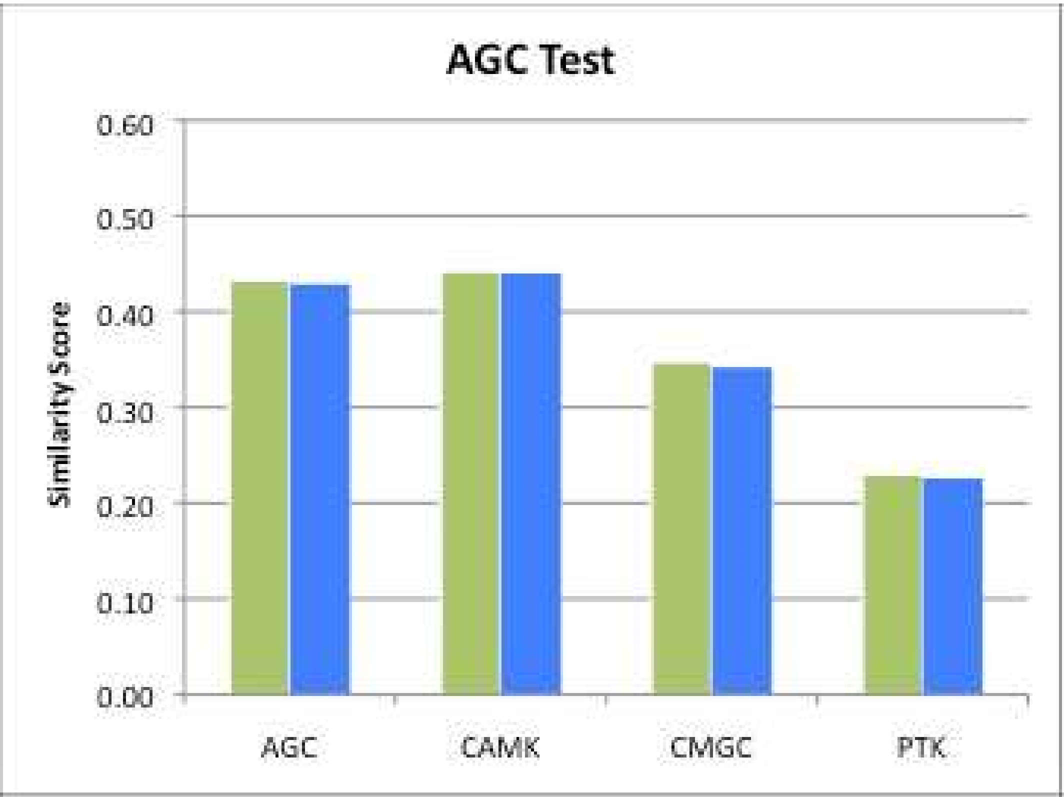
Table 11 Amino acid hyrophobicity index matrix. Amino acid order: ARNDCQEGHILKMFPSTWYV [21].

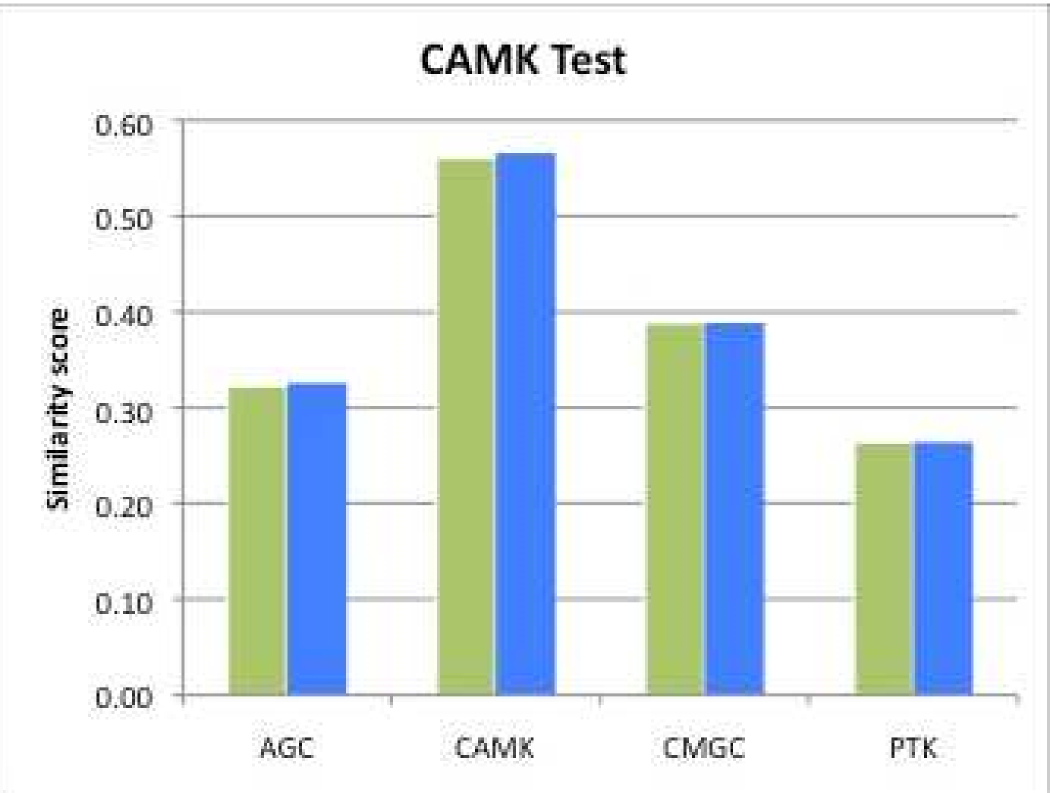

Projection of the 3D protein structure to obtain 2D starting coordinates
Calculation of error between residue distances in 2D and 3D
Monte Carlo Acceptance Criteria
Rejected
Accepted
Generated new 2D coordinates
Write 2D coordinate file

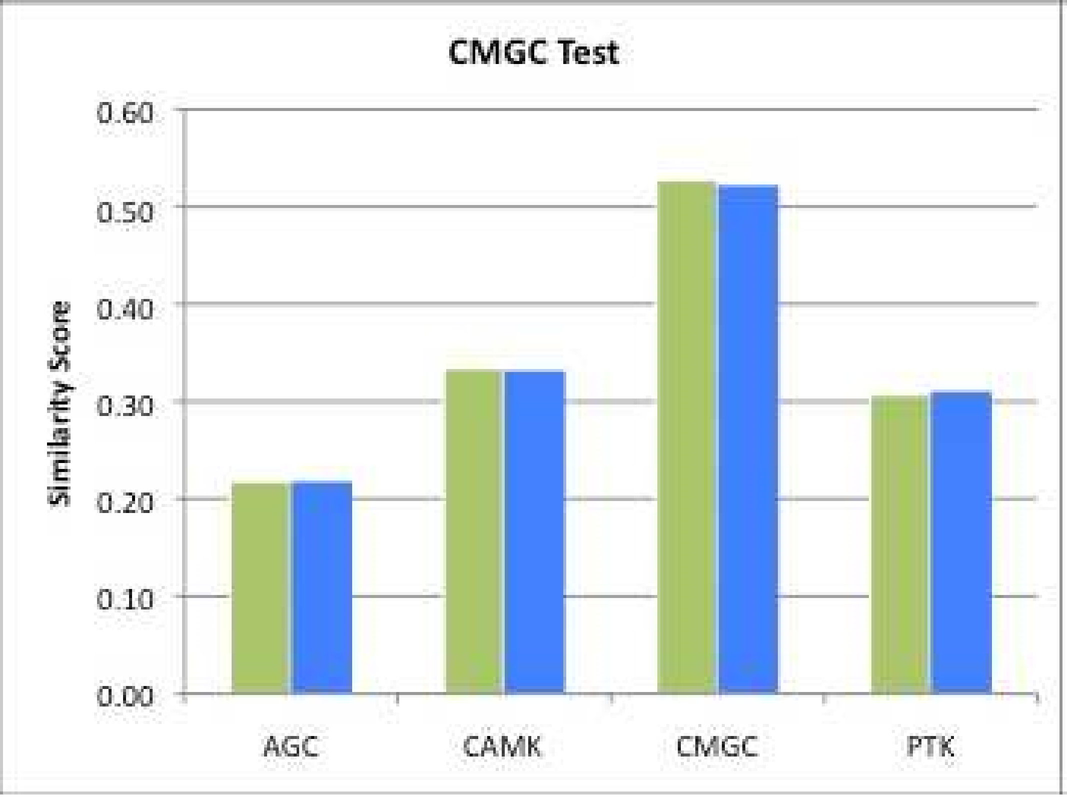


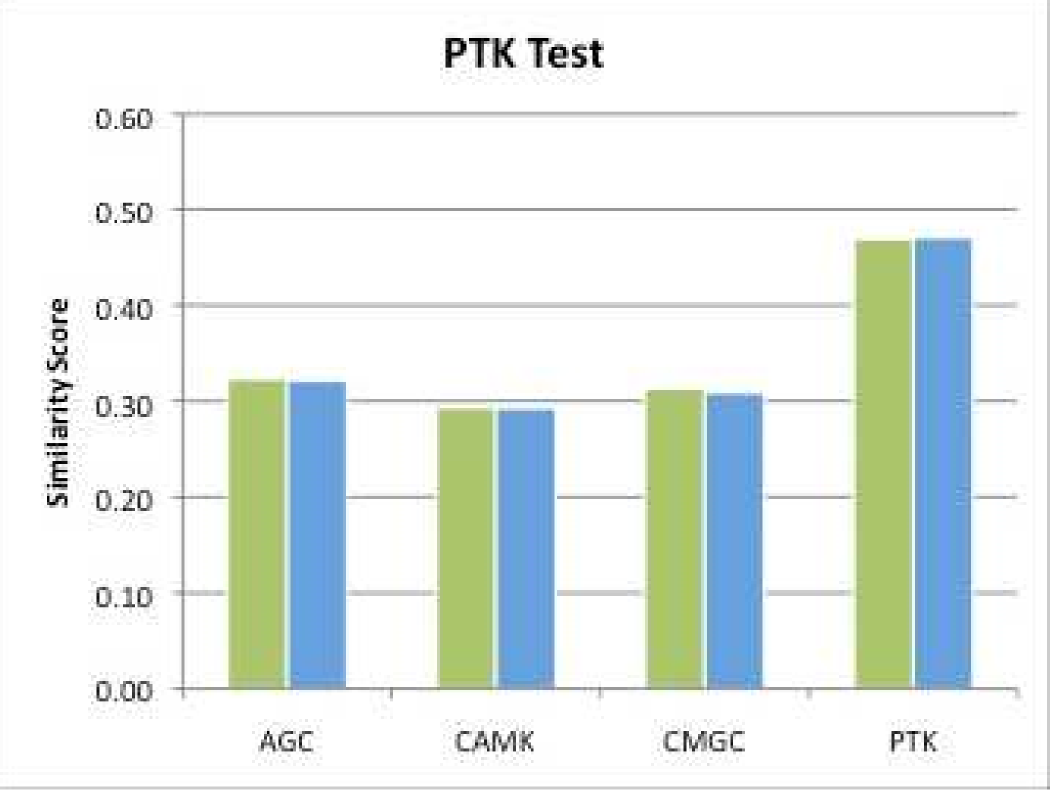

Initial projection
20
15
10
5
0
-5
-10
-15
-20
-20
-15
-10
-5
0
5
10
15
20
25

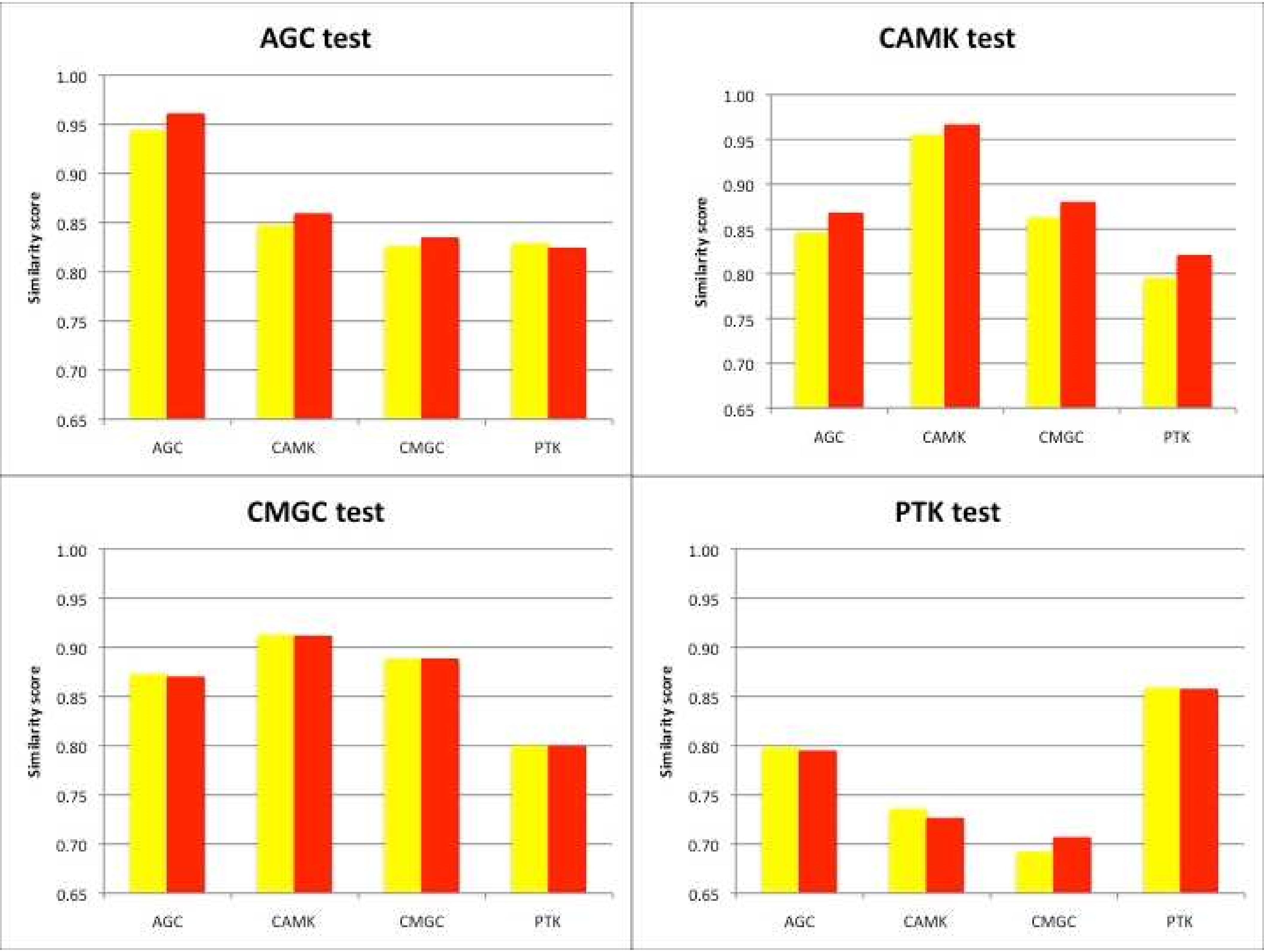

final map
1
0.5
0
-0.5
-1
-1
-0.5
0
0.5
1

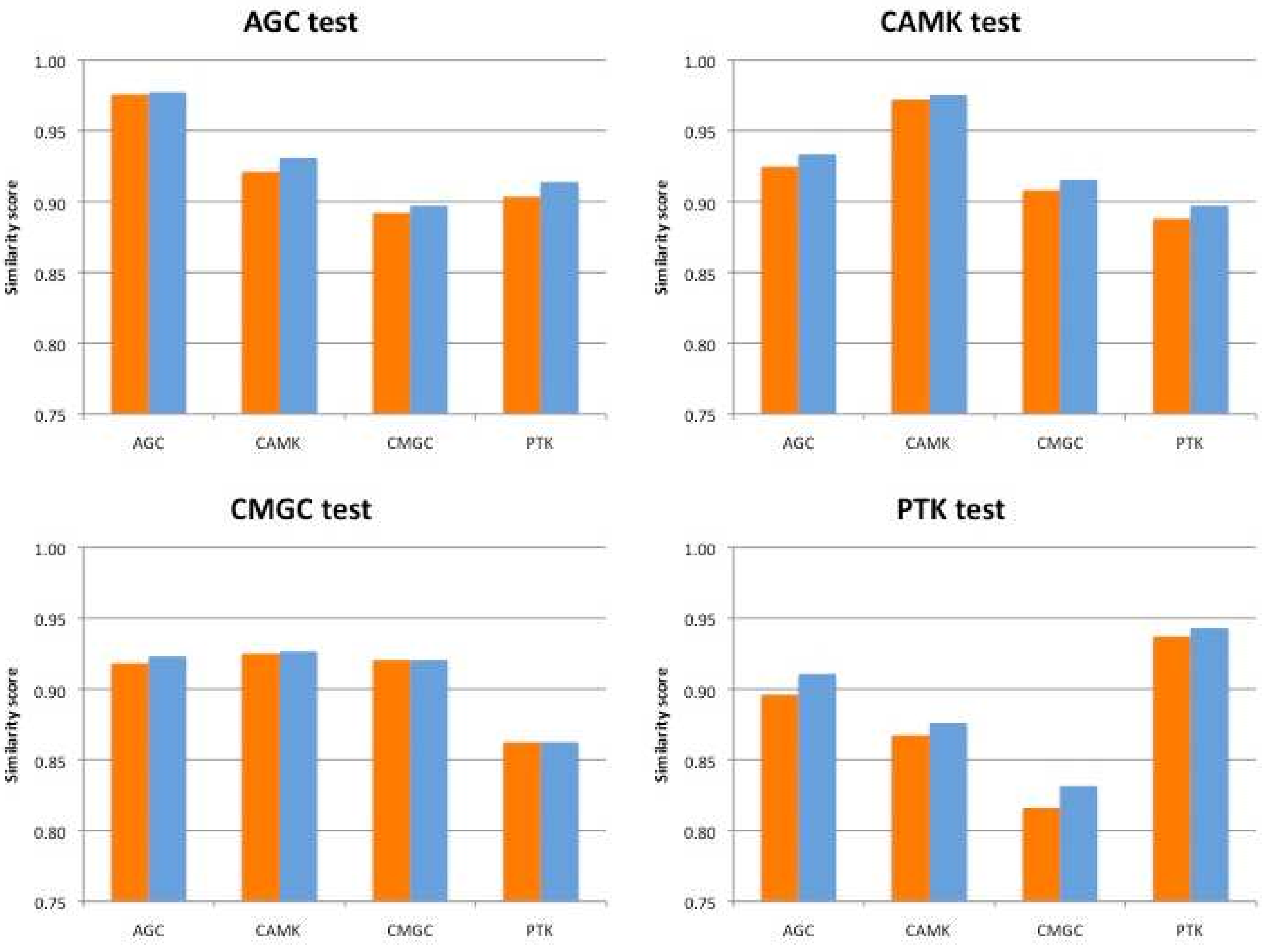

AGC Test
Similarity Score
0.60
0.50
0.40
0.30
0.20
0.10
0.00
AGC
CAMK
CMGC
PTK

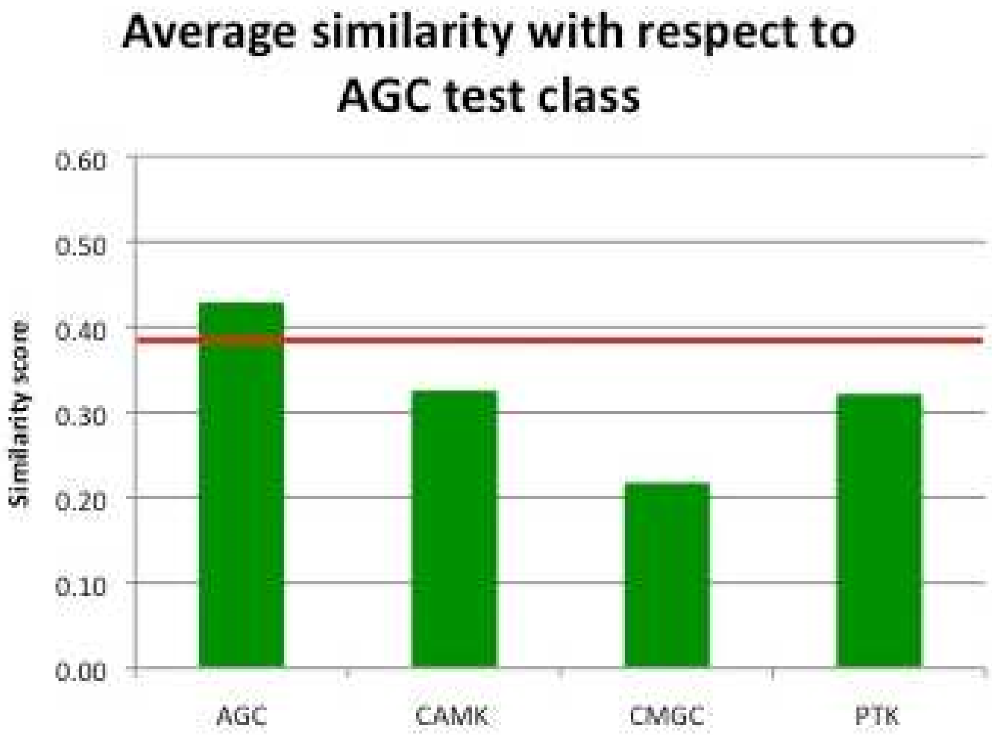

CAMK Test
Similarity score
0.60
0.50
0.40
0.30
0.20
0.10
0.00
AGC
CAMK
CMGC
PTK

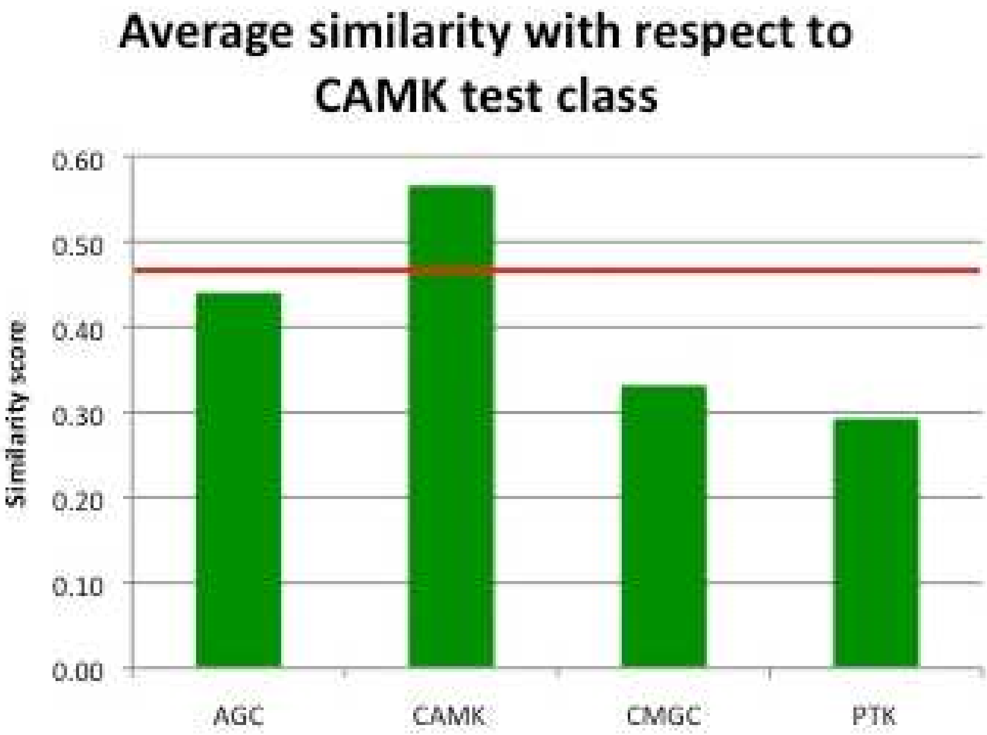

CMGC Test
Similarity Score
0.60
0.50
0.40
0.30
0.20
0.10
0.00
AGC
CAMK
CMGC
PTK

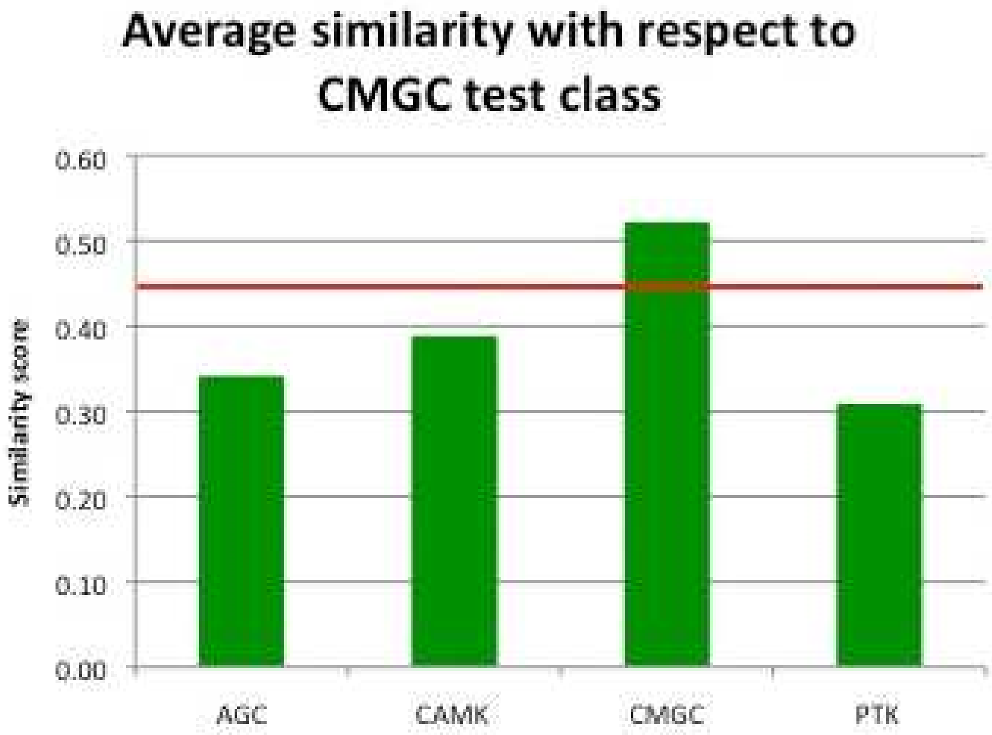

PTK Test
Similarity Score
0.60
0.50
0.40
0.30
0.20
0.10
0.00
AGC
CAMK
CMGC
PTK

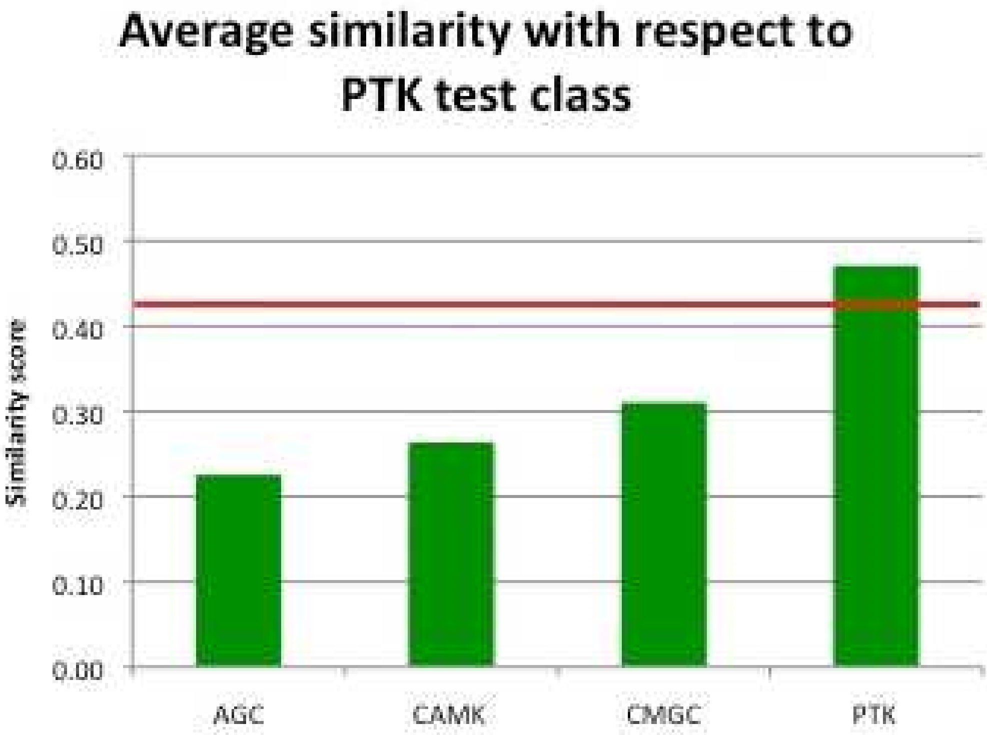

AGC test
CAMK test
CMGC test
PTK test
Similarity score
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
AGC
CAMK
CMGC
PTK

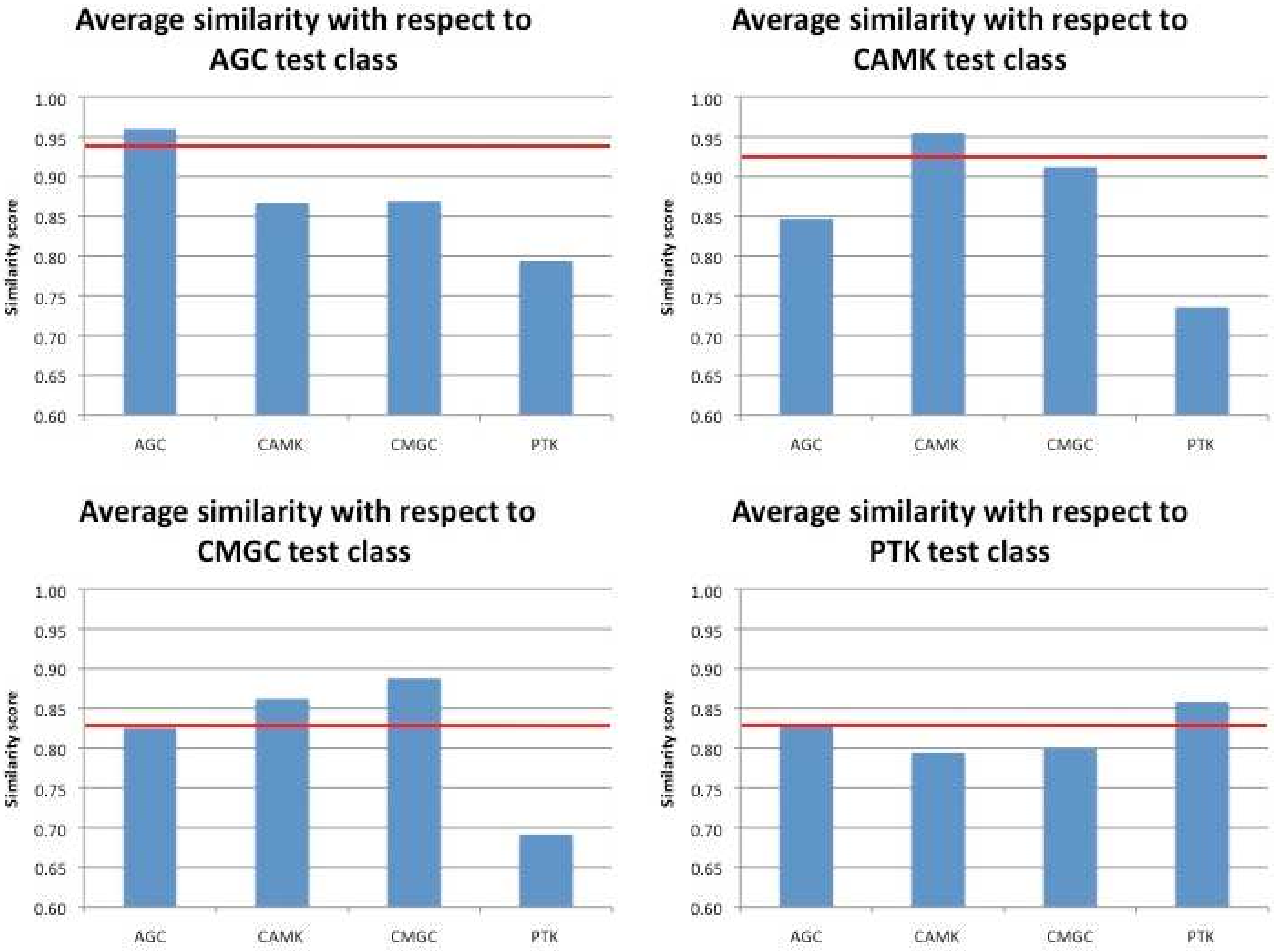

AGC test
CAMK test
CMGC test
PTK test
Similarity score
1.00
0.95
0.90
0.85
0.80
0.75
AGC
CAMK
CMGC
PTK


Average similarity with respect to AGC test class
Similarity score
0.60
0.50
0.40
0.30
0.20
0.10
0.00
AGC
CAMK
CMGC
PTK


Average similarity with respect to CAMK test class
Similarity score
0.60
0.50
0.40
0.30
0.20
0.10
0.00
AGC
CAMK
CMGC
PTK


Average similarity with respect to CMGC test class
Similarity score
0.60
0.50
0.40
0.30
0.20
0.10
0.00
AGC
CAMK
CMGC
PTK


Average similarity with respect to PTK test class
Similarity score
0.60
0.50
0.40
0.30
0.20
0.10
0.00
AGC
CAMK
CMGC
PTK

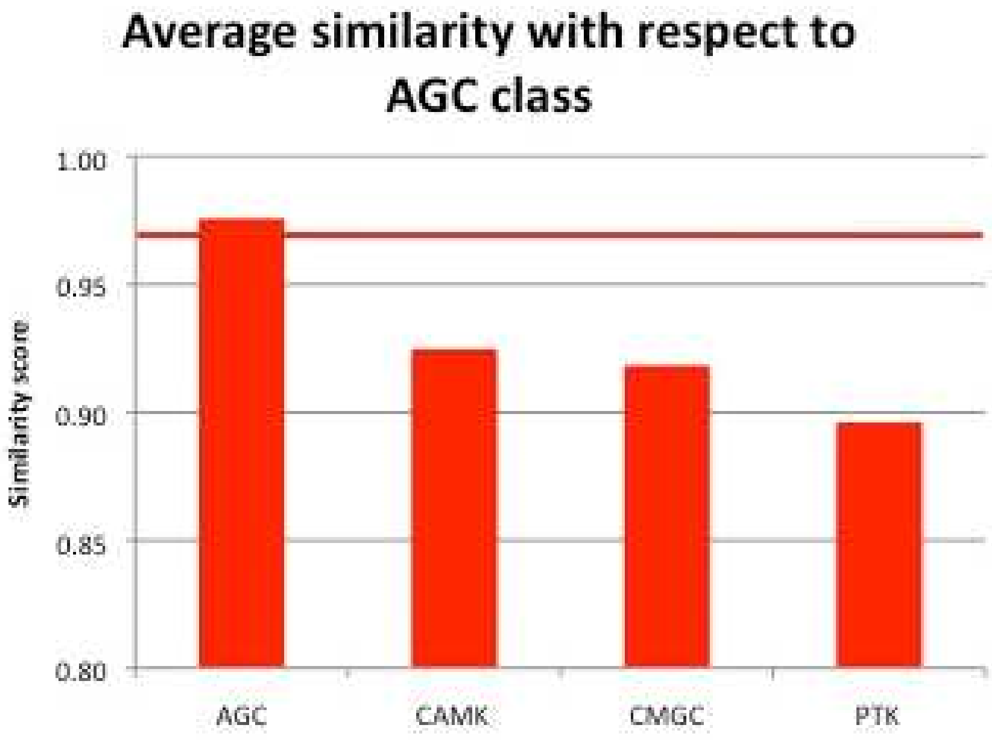


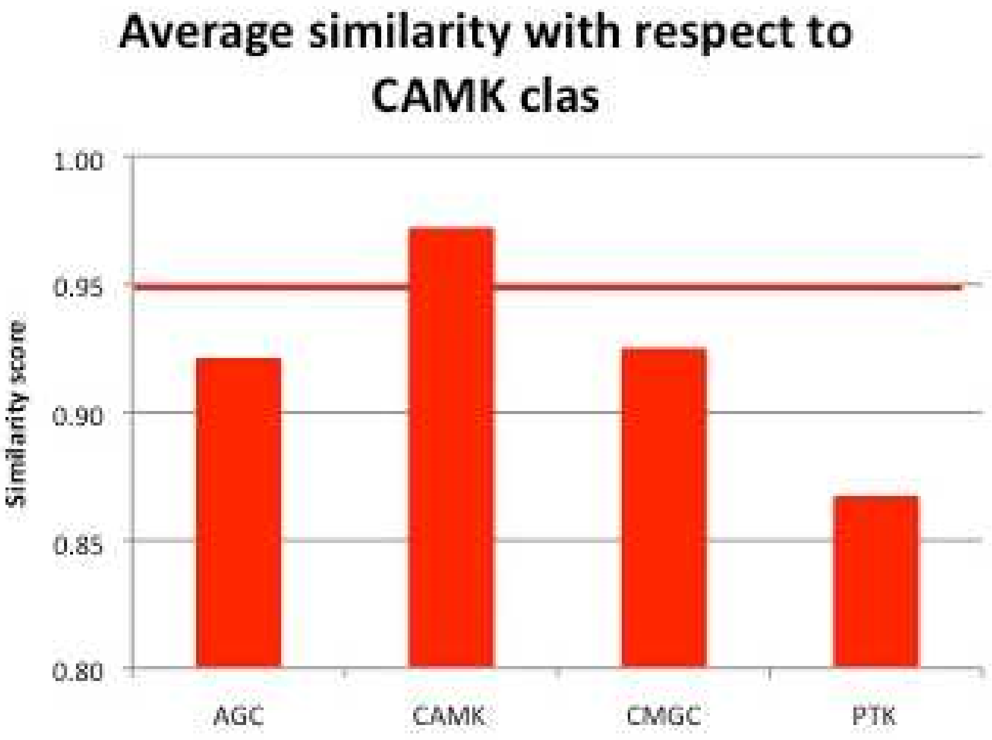
Average similarity with respect to AGC test class

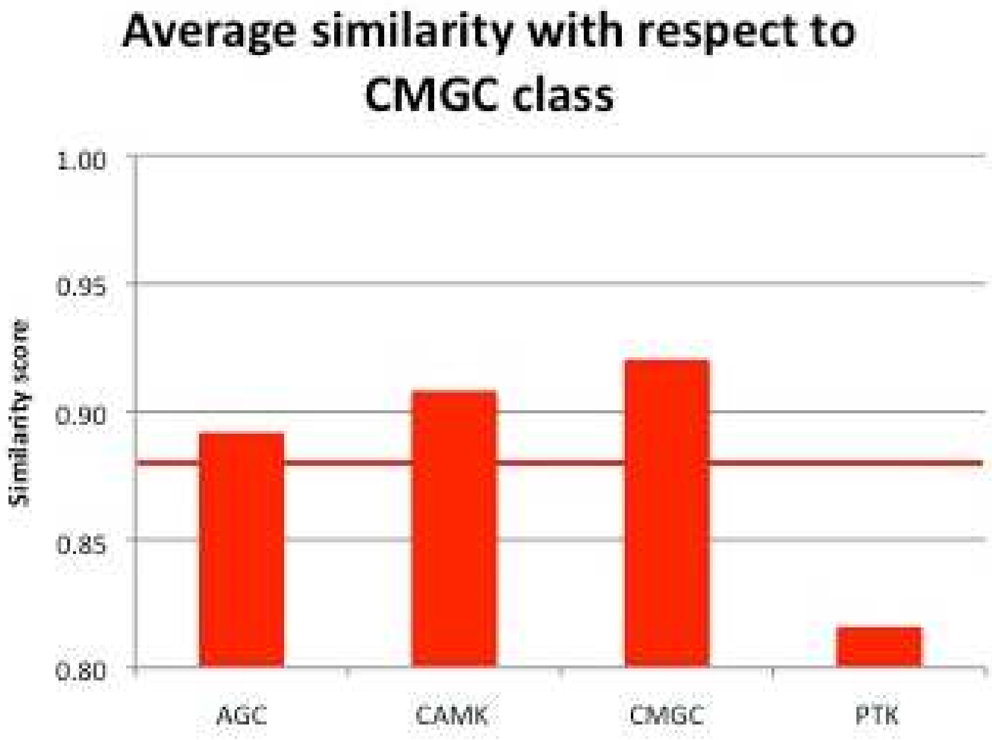
Average similarity with respect to CAMK test class

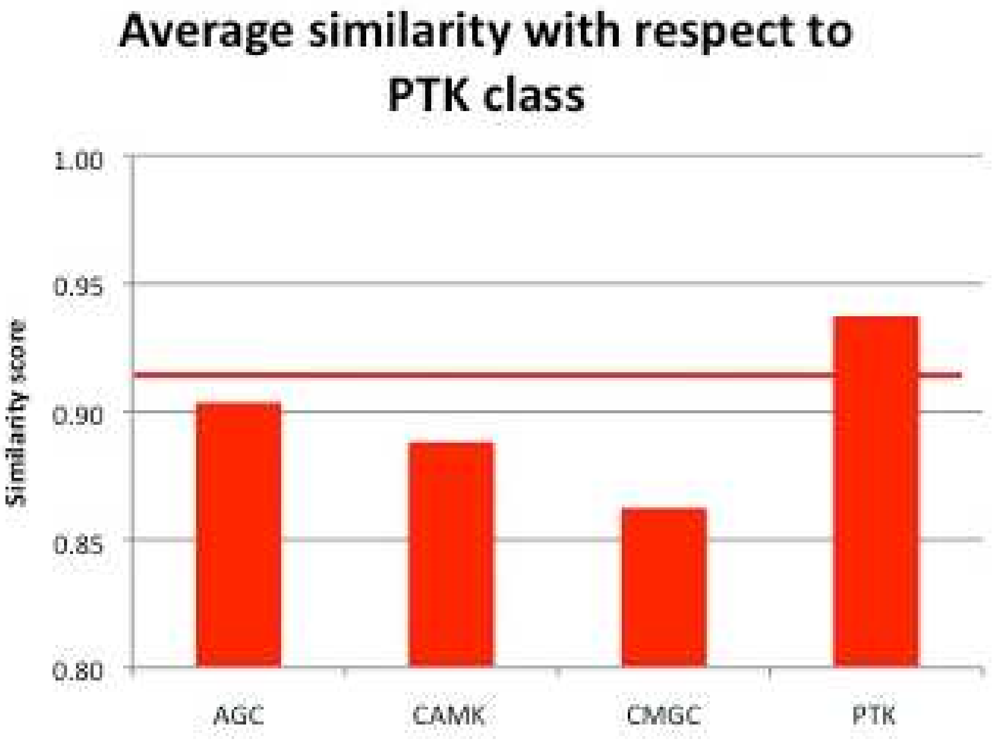
Average similarity with respect to CMGC test class

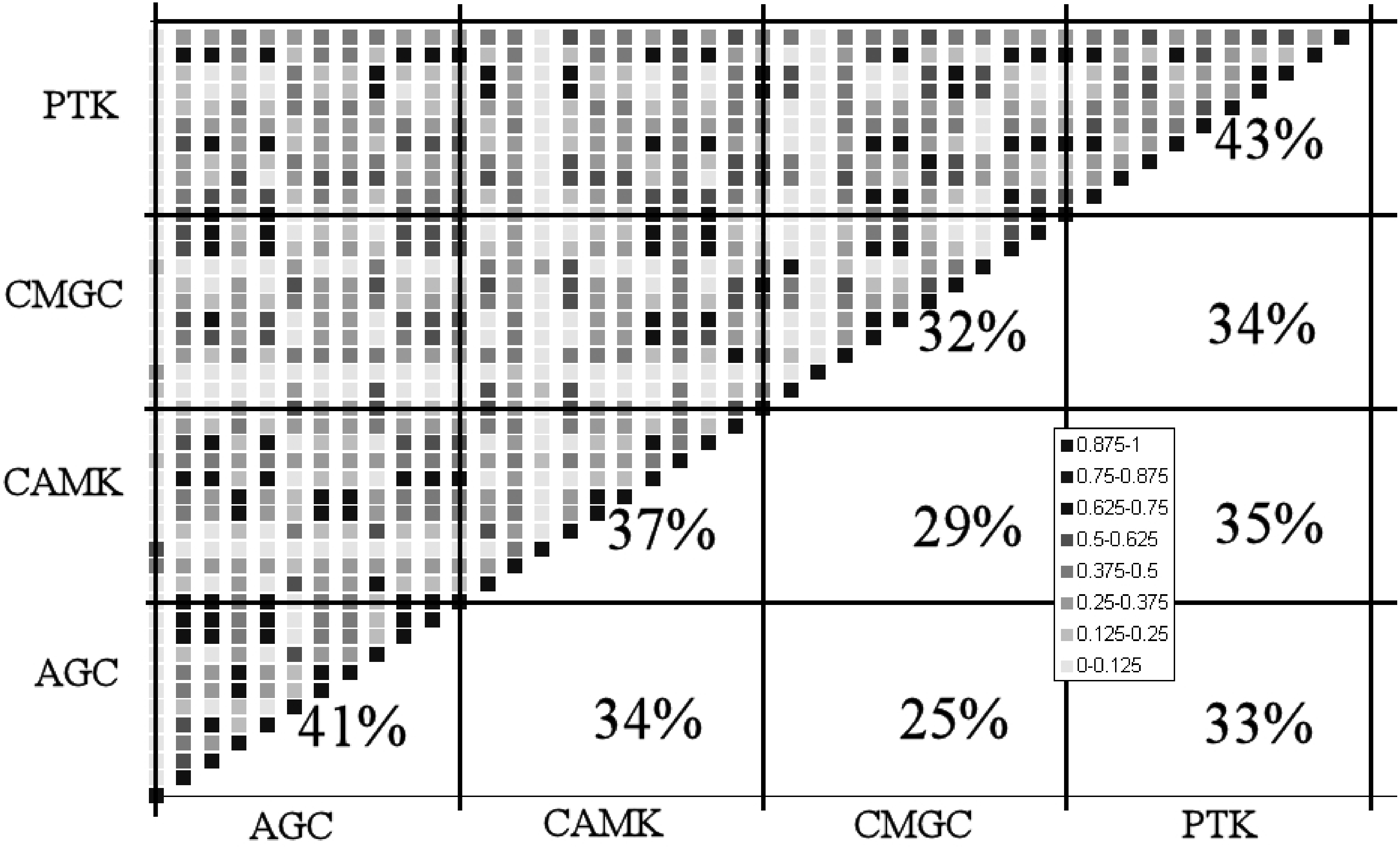
Average similarity with respect to PTK test class

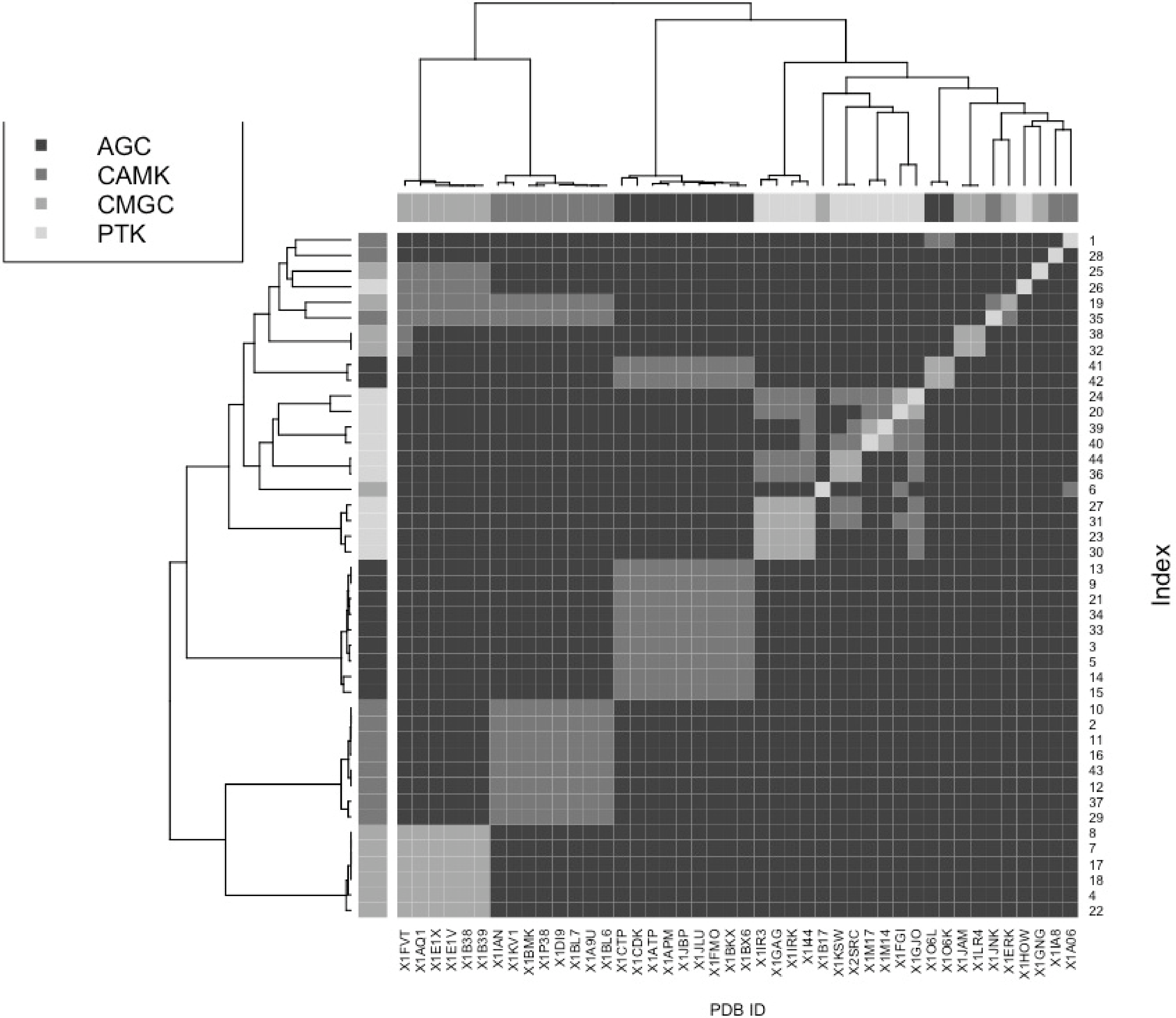

Average similarity with respect to AGC class

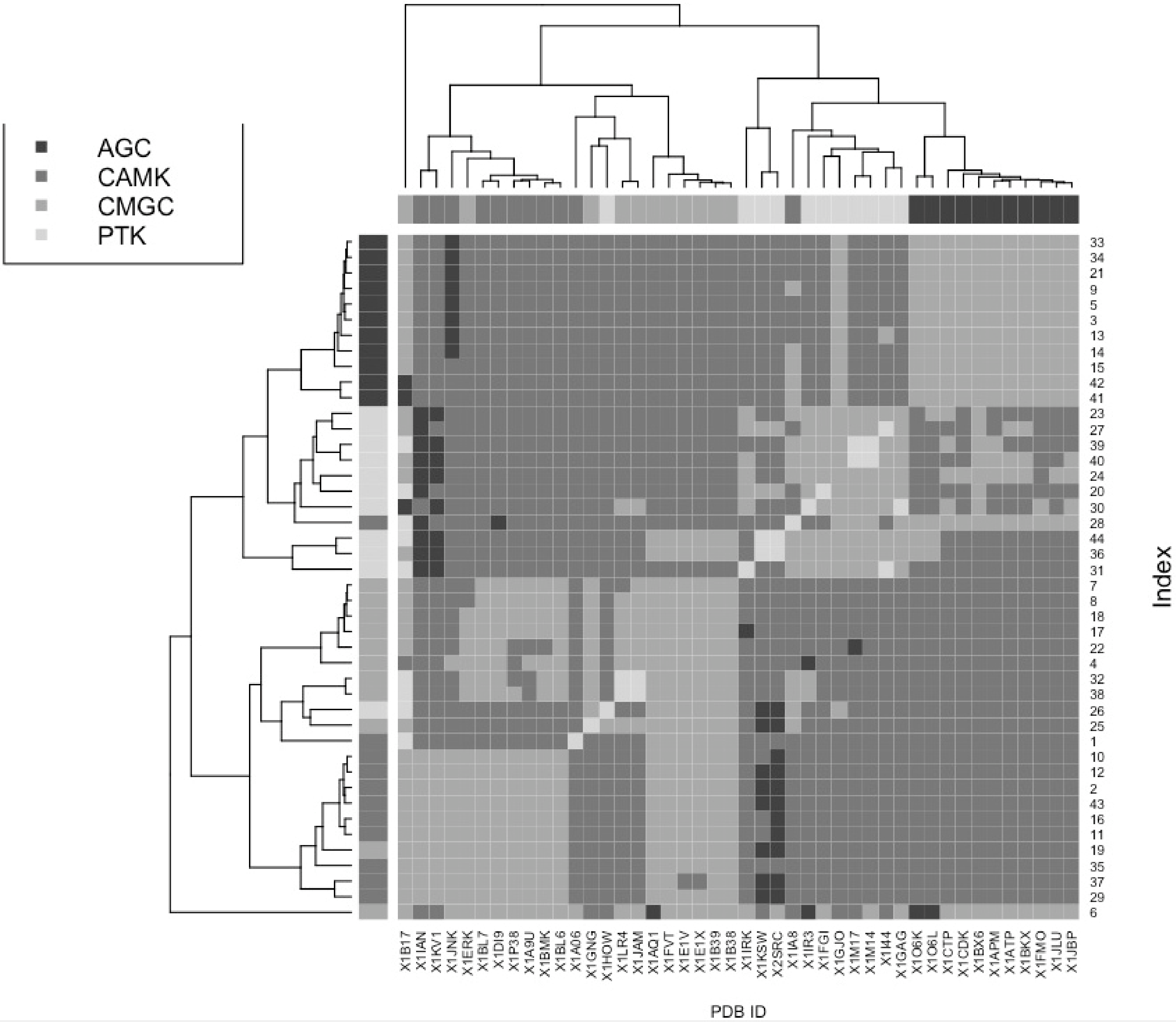

Average similarity with respect to CAMK clas

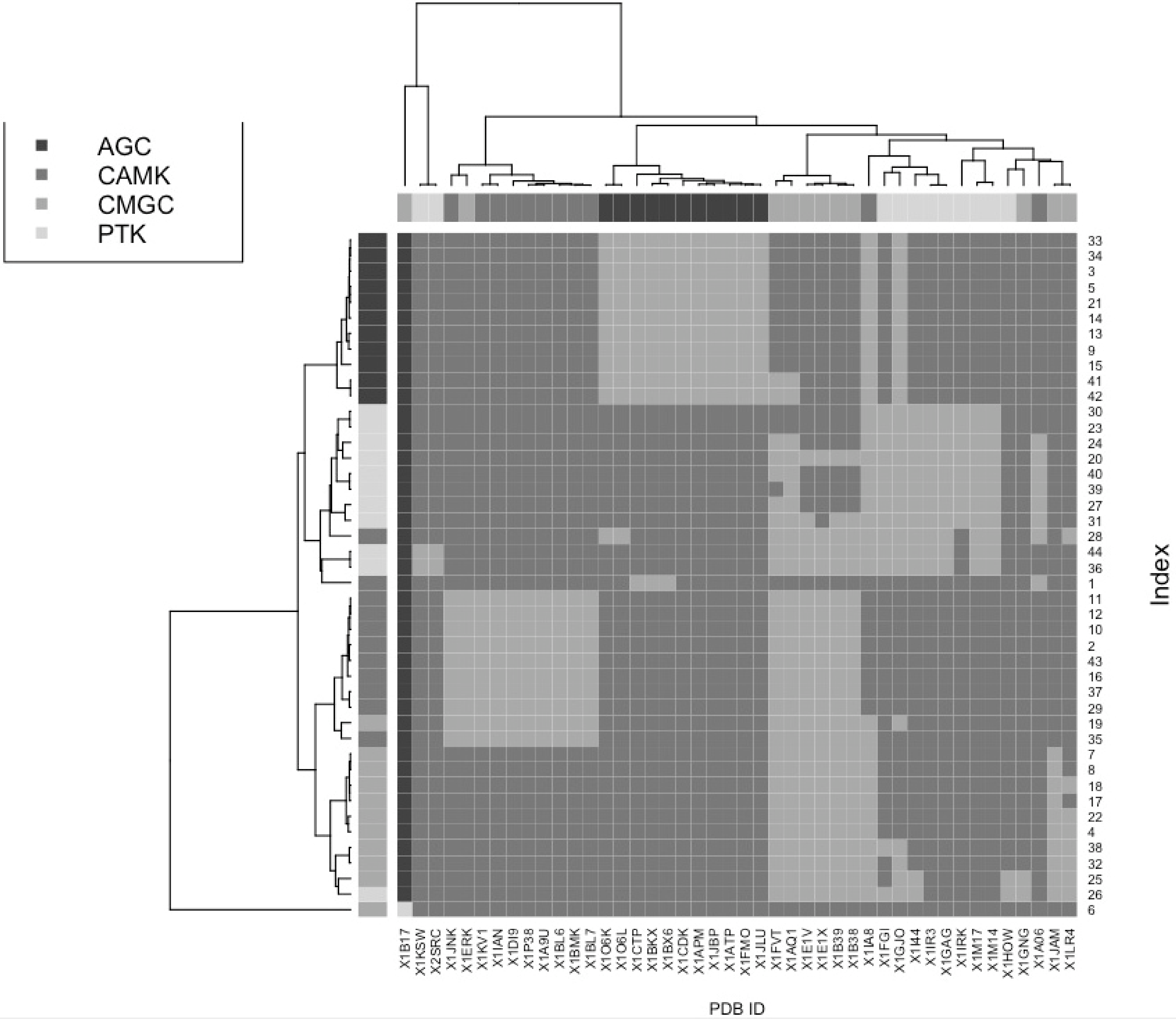

Average similarity with respect to CMGC class

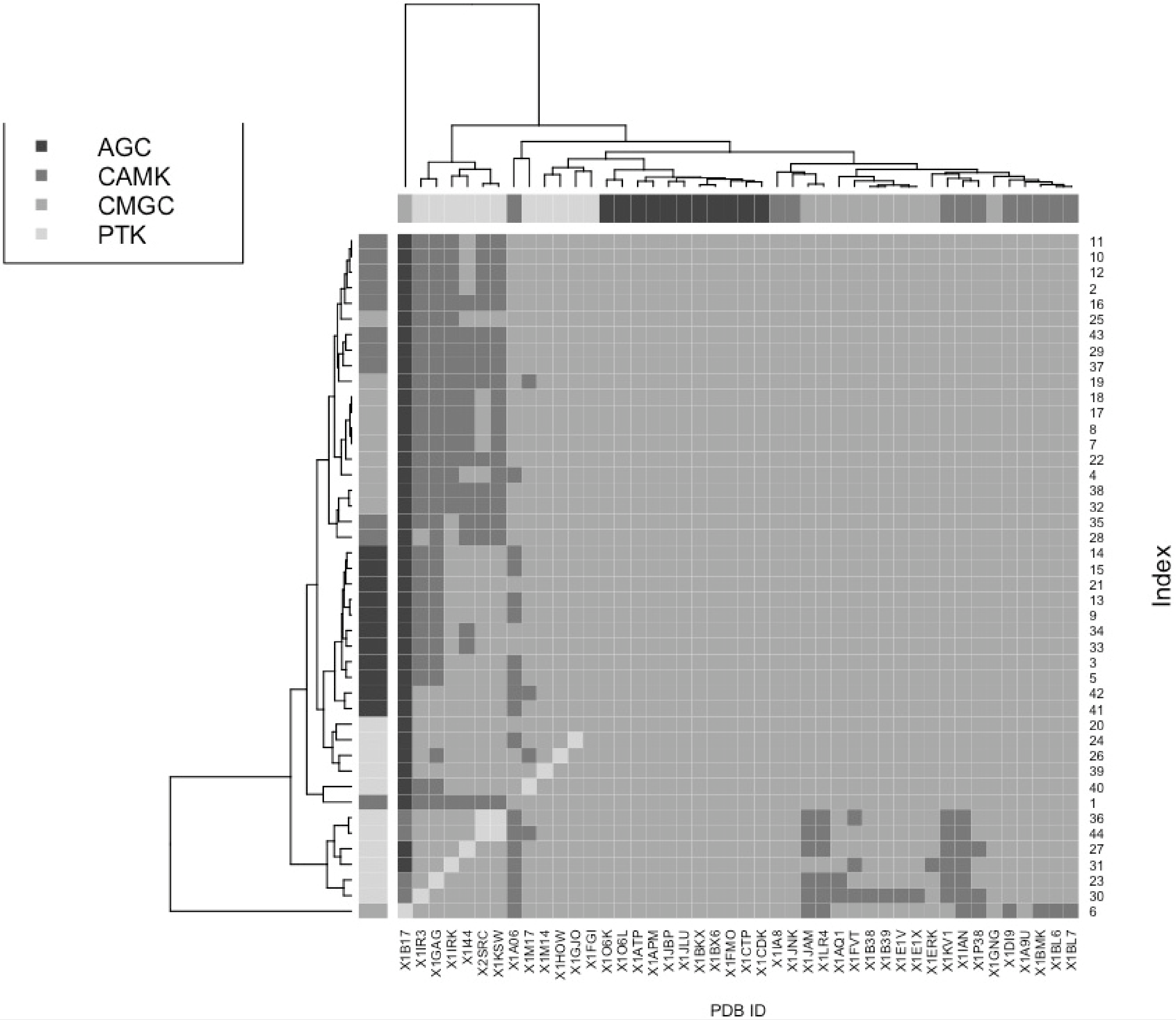

Average similarity with respect to PTK class


PTK
CMGC
CAMK
AGC
43%
32%
34%
37%
29%
35%
41%
34%
25%
33%
0.875-1
0.75-0.875
0.625-0.75
0.5-0.625
0.375-0.5
0.25-0.375
0.125-0.25
0-0.125
AGC
CAMK
CMGC
PTK


AGC
CAMK
CMGC
PTK
X1FVT
X1AQ1
X1E1X
X1E1V
X1B38
X1B39
X1IAN
X1KV1
X1BMK
X1P38
X1DI9
X1BL7
X1A9U
X1BL6
X1CTP
X1CDK
X1ATP
X1APM
X1JBP
X1JLU
X1FMO
X1BKX
X1BX6
X1IR3
X1GAG
X1IRK
X1I44
X1B17
X1KSW
X2SRC
X1M17
X1M14
X1FGI
X1GJO
X1O6L
X1O6K
X1JAM
X1LR4
X1JNK
X1ERK
X1HOW
X1GNG
X1IA8
X1A06
PDB ID
1
28
25
26
19
35
38
32
41
42
24
20
39
40
44
36
6
27
31
23
30
13
9
21
34
33
3
5
14
15
10
2
11
16
43
12
37
29
8
7
17
18
4
22
Index


AGC
CAMK
CMGC
PTK
X1B17
X1IAN
X1KV1
X1JNK
X1ERK
X1BL7
X1DI9
X1P38
X1A9U
X1BMK
X1BL6
X1A06
X1GNG
X1HOW
X1LR4
X1JAM
X1AQ1
X1FVT
X1E1V
X1E1X
X1B39
X1B38
X1IRK
X1KSW
X2SRC
X1IA8
X1IR3
X1FGI
X1GJO
X1M17
X1M14
X1I44
X1GAG
X1O6K
X1O6L
X1CTP
X1CDK
X1BX6
X1APM
X1ATP
X1BKX
X1FMO
X1JLU
X1JBP
PDB ID
33
34
21
9
5
3
13
14
15
42
41
23
27
39
40
24
20
30
28
44
36
31
7
8
18
17
22
4
32
38
26
25
1
10
12
2
43
16
11
19
35
37
29
6
Index


AGC
CAMK
CMGC
PTK
X1B17
X1KSW
X2SRC
X1JNK
X1ERK
X1KV1
X1IAN
X1DI9
X1P38
X1A9U
X1BL6
X1BMK
X1BL7
X1O6K
X1O6L
X1CTP
X1BKX
X1BX6
X1CDK
X1APM
X1JBP
X1ATP
X1FMO
X1JLU
X1FVT
X1AQ1
X1E1V
X1E1X
X1B39
X1B38
X1IA8
X1FGI
X1GJO
X1I44
X1IR3
X1GAG
X1IRK
X1M17
X1M14
X1HOW
X1GNG
X1A06
X1JAM
X1LR4
PDB ID
33
34
3
5
21
14
13
9
15
41
42
30
23
24
20
40
39
27
31
28
44
36
1
11
12
10
2
43
16
37
29
19
35
7
8
18
17
22
4
38
32
25
26
6
Index


AGC
CAMK
CMGC
PTK
11
10
12
2
16
25
43
29
37
19
18
17
8
7
22
4
38
32
35
28
14
15
21
13
9
34
33
3
5
42
41
20
24
26
39
40
1
36
44
27
31
23
30
6
Index
X1B17
X1IR3
X1GAG
X1IRK
X1I44
X2SRC
X1KSW
X1A06
X1M17
X1M14
X1HOW
X1GJO
X1FGI
X1O6K
X1O6L
X1ATP
X1APM
X1JBP
X1JLU
X1BKX
X1BX6
X1FMO
X1CTP
X1CDK
X1IA8
X1JNK
X1JAM
X1LR4
X1AQ1
X1FVT
X1B38
X1B39
X1E1V
X1E1X
X1ERK
X1KV1
X1IAN
X1P38
X1GNG
X1DI9
X1A9U
X1BMK
X1BL6
X1BL7
PDB ID