Measuring the Similarity of Protein Structures Using Image Local Feature Descriptors SIFT and SURF

Morihiro Hayashida
Bioinformatics Center
Institute for Chemical Research
Kyoto University
Kyoto 611–0011, Japan
Email: morihiro@kuicr.kyoto-u.ac.jp

Hitoshi Koyano
Department of Data Science
Institute for Advancement of Clinical and Translational Science
Graduate School of Medicine
Kyoto University
Kyoto 606–8507, Japan

Tatsuya Akutsu
Bioinformatics Center
Institute for Chemical Research
Kyoto University
Kyoto 611–0011, Japan

***Abstract*—Understanding of protein structures is important to find their functions. Many methods such as structural alignment, alignment-free similarity, and use of structural fragments have been developed for finding similar protein structures. In our previous study, we transformed protein structures into images each pixel of which represents the distance between the corresponding C_α atoms, and proposed similarity measures between two protein structures based on Kolmogorov complexity using image compression algorithms.**

In this paper, we examine efficient and effective image recognition techniques, SIFT (Scale Invariant Feature Transform) and SURF (Speeded Up Robust Features), which are invariant to image scaling, translation, and rotation, and partially invariant to affine or three-dimensional projection. We propose similarity based on SIFT and SURF, and apply it to classification of several protein structures. The results suggest that the similarity based on SURF outperforms several existing similarity measures including the compression-based similarity measures in our previous study, and that SIFT and SURF are useful for recognizing protein structures as well as objects in images.

I. Introduction

Finding similar protein structures is one of important tasks for understanding protein functions and evolution. Several databases such as CATH [1] and SCOP [2] have been developed for storing classified protein structures. Since evolutionary selection pressure is exerted on structures, structures are more conserved than sequences. Many structural alignment methods have been developed, which include DALI [3], Stralign [4], combinatorial extension (CE) of aligned fragment pairs [5], iterative double dynamic programming [6], secondary-structure matching (SSM) [7], TM-align [8], and MICAN [9]. These methods take expensive computational time, and are not suitable for aligning a protein structure with all structures in the whole database of PDB (Protein Data Bank) [10].

Several methods without aligning structures have been developed for measuring similarity between protein structures. Krasnogor and Pelta [11] transformed distances between C_α atoms, that is, the contact map as also in [12], into a 0-1 matrix, which was further transformed into a 0-1 sequence. They approximated Kolmogorov complexities by compression ratios of 0-1 sequences obtained from two protein structures, and calculated Universal Similarity Metric (USM) [13]. In our previous study [14], we transformed contact maps into two-dimensional images, applied image compression algorithms, JPEG, GIF, PNG, IFS [15], SPC [16], and MSPC [14], to such images, and calculated NCD (Normalized Compression Dissimilarity) and ACD (Anticommutative Compression Dissimilarity), which were derived from USM. We applied some similarity measures to classification of protein structures, calculated ROC (Receiver Operating Characteristic) curves and AUC (Area Under ROC Curve), and reported that the AUC values by NCD and ACD using the MSPC image compression algorithm were larger than those by some existing methods including the method by Krasnogor and Pelta [11]. It should be noted that Rocha et al. insisted that USM was not suitable for discriminating protein domains from their computational experiments [17].

There are other efficient approaches to measuring the similarity between protein structures. Carugo and Pongor proposed a probability of identity (PRIDE score), which is calculated from histograms of distances between C_α atoms, according to the idea that the distribution of distances between a C_α atom and that separated by 3 to 30 residues indicates certain characteristics concerning protein folds [18]. A protein backbone is represented as a curve, and how two curves interact with each other can be measured by the Gauss integral in knot theory. According to this theory, Røgen and Fain proposed Scaled Gauss Metric (SGM) [19]. Choi et al. subdivided the distance matrix into submatrices describing secondary and tertiary features, extracted several representative medoid submatrices, and constructed a vector, called LFF (Local Feature Frequency) profile vector, containing the frequencies of occurrences of these submatrices from a protein structure [20]. Then, the similarity was calculated as the Euclidean or cosine distance between two corresponding LFF profile vectors. Zotenko et al. used a triplet of secondary structure elements (SSE) as a structural fragment, and represented a protein structure by the spatial conformation of SSEs [21]. Lo et al. developed *i*SARST server [22] for efficient protein structural similarity searches by utilizing Ramachandran sequential transformation [23]. Budowski-Tal et al. proposed FragBag that represents a protein structure as a bag of words (BoW) of backbone fragments [24], where BoW is collections of local features, and a document can be represented by the frequencies of occurrences of the words in the BoW. Molloy et al. proposed an alternative

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

representation of a protein structure using FragBag based on Latent Dirichlet Allocation (LDA) probabilistic model to find latent groups [25]. These methods represent a protein structure as a vector, and try to find similar vectors.

On the other hand, in the fields of image processing and computer vision, efficient and effective methods have been developed for image recognition. Lowe proposed an image feature generation method SIFT (Scale Invariant Feature Transform), which transforms an image into local feature vectors that are invariant to image translation, scaling, and rotation, and partially invariant to affine or three-dimensional projection [26], [27]. Instead of all locations in an image, some characteristic key points are selected, and are transformed into local feature vectors, respectively. Bay et al. approximated SIFT, and proposed a faster method SURF (Speeded Up Robust Features) [28]. In this paper, we transform distance matrices between C_α atoms into images, propose similarity based on SIFT and SURF, and apply it to classification of several proteins. The results suggest that the similarity based on SURF outperforms several existing similarity measures including the compression-based similarity measures in our previous study, and that SIFT and SURF are useful for recognizing protein structures as well as objects in images.

II. Methods

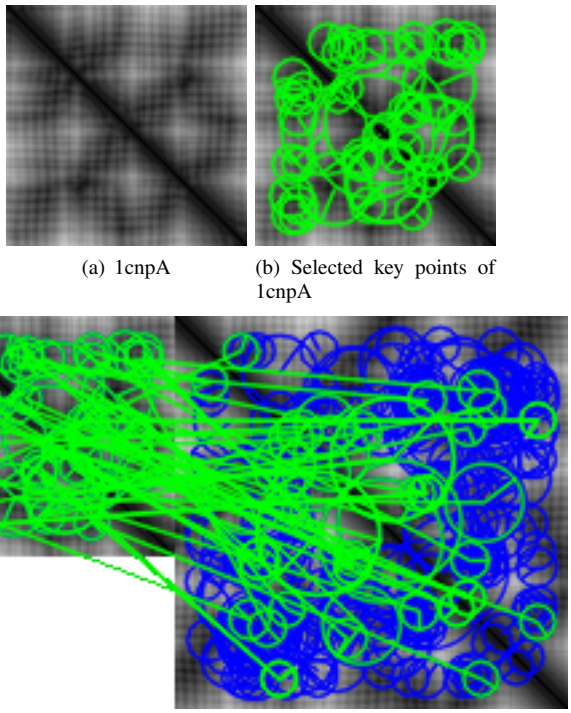
We briefly review the image local feature descriptors SIFT [26], [27] and SURF [28], and propose the similarity between protein structures based on SIFT and SURF.

A. SIFT (Scale Invariant Feature Transform)

SIFT generates several feature vectors from an image, and consists of two parts, selection of stable key points, and construction of feature vectors. In the selection of key points, candidates of key points are identified as maxima or minima in a difference-of-Gaussian function convolved with the image. Each candidate is fit to a three-dimensional quadratic function, the interpolated point of the extrema is determined, and all extrema with the absolute value of the quadratic function less than 0.03 are eliminated, that is, key points having low contrast are eliminated. In addition, candidates of key points on edges of image objects are eliminated using the principal curvature of the difference-of-Gaussian function.

In the construction of a feature vector for each key point, first a local orientation is assigned to the key point. An orientation histogram, which has 36 bins covering the 360 degree range, is constructed from the gradient magnitude. The orientation with the highest peak in the histogram is selected, and the coordinate of the original image is rotated by the selected orientation for obtaining the feature vector. The region around the key point is divided into 4×4 blocks. In each block, gradient magnitudes and an orientation histogram with 8 bins are calculated. Then, the SIFT descriptor at the key point consists of 4×4×8=128 entries of the orientation histograms.

Figure 1 illustrates image local feature descriptors SIFT and SURF, where (a) shows an input grayscale image transformed from a protein structure with PDB code 1cnpA, and (b) shows detected key points with scale and orientation for the image of (a).



(a) 1cnpA (b) Selected key points of 1cnpA

(c) Matching between key points of 1cnpA and 1ash

Fig. 1. Illustration on image local feature descriptors SIFT and SURF. (a) Input grayscale image transformed from a protein structure with PDB code 1cnpA. (b) Detected key points (denoted by the center of each circle) with scale (denoted by the radius of the circle) and orientation (denoted by the line in the circle) for the image of 1cnpA. (c) Matching between key points of images of 1cnpA and 1ash, denoted by lines.

B. SURF (Speeded Up Robust Features)

In order to detect key points, SURF uses a Hessian-matrix approximation instead of the difference-of-Gaussian function of SIFT. For that purpose, they made use of integral images each pixel of which has the sum of all pixels within the region formed by the origin and the pixel itself. The sum of all pixels within any rectangular region can be calculated from the integral image in constant time. For the extraction of the descriptor, the region around the key point is divided into 4×4 blocks, and Haar wavelet responses are summed up over each sub-region. Let w_h and w_v be the responses in horizontal and vertical directions, respectively. Then, the SURF descriptor at the key point consists of $\sum w_h$, $\sum w_v$, $\sum |w_h|$, and $\sum |w_v|$ for each sub-region.

C. Similarity Based on SIFT and SURF

We propose the similarity between protein structures based on SIFT and SURF. First, we transform a protein structure into the grayscale image each pixel of which has a value $\lfloor c \cdot d_{ij} \rfloor$ because the value of a pixel must be an integer, where c is a positive constant value, d_{ij} denotes the distance between C_α atoms of i-th and j-th residues, and $\lfloor x \rfloor$ denotes the floor function of x, which gives the largest integer less than or equal to x.

Let $\boldsymbol{f}_i^{(p)}$ be the feature vector at i-th key point calculated using SIFT or SURF for protein p. Then, we propose the

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

TABLE I. THE CHEW-KEDEM DATASET. IT CONTAINS 19 MAINLY ALPHA PROTEINS, 7 MAINLY BETA PROTEINS, AND 10 ALPHA-BETA PROTEINS.

Class of CATH	Superfamily	Proteins
Mainly alpha	Globins (1.10.490.10)	1ash, 1babA, 1babB, 1eca, 1flp, 1hlb, 1hlm, 1ithA, 1lh2, 1mba, 1myt, 2hbg, 2lhb, 2vhb, 2vhbA, 3sdhA, 5mbn
	EF-hand (1.10.238.10)	1cnpA
	Trp operon repressor (1.10.1270.10)	1jhgA
Mainly beta	Immunoglobulins (2.60.40.10)	1cd8, 1cdb, 1ci5A, 1hnf01, 1neu, 1qa9A, 1qfoA
Alpha-beta	Enolase-like, N-terminal domain (3.30.390.10)	1chrA1, 2mnr01, 4enl01
	P-loop containing nucleotide triphosphate hydrolases (3.40.50.300)	1aa9, 1gnp, 1qraA, 5p21, 6q21A
	Glutamine phosphoribosyl-pyrophosphate (3.60.20.10)	1ct9A1
	Divalent-metal-dependent TIM barrel enzymes (3.20.20.150)	6xia

similarity between two proteins p and q defined by

$$\max\Big\{\frac{1}{N_p}\sum_{i=1}^{N_p}\min_{j=1,\ldots,N_q}|\boldsymbol{f}_i^{(p)}-\boldsymbol{f}_j^{(q)}|,\ \frac{1}{N_q}\sum_{j=1}^{N_q}\min_{i=1,\ldots,N_p}|\boldsymbol{f}_j^{(q)}-\boldsymbol{f}_i^{(p)}|\Big\}, \quad (1)$$

where N_p denotes the number of key points for the corresponding image to protein p, and $|\boldsymbol{f}| = \sqrt{(\boldsymbol{f},\boldsymbol{f})}$. Figure 1(c) shows an example of matching between key points of images for proteins of PDB codes 1cnpA and 1ash, where two key points in different images are connected by a line if the corresponding feature vector in the image of 1ash is closest to that in the image of 1cnpA.

III. COMPUTATIONAL EXPERIMENTS

A. Data and Implementation

For evaluation of our proposed similarity between protein structures, we used two datasets, the Chew-Kedem dataset [29] and the Sierk-Pearson dataset [30], which were also used in [11], [14], [31]. The Chew-Kedem dataset consists of 36 proteins identified by PDB codes including 19 mainly alpha proteins, 7 mainly beta proteins, and 10 alpha-beta proteins as shown in Table I. The Sierk-Pearson dataset consists of 86 proteins including 20 mainly alpha proteins, 26 mainly beta proteins, and 40 alpha-beta proteins as shown in Table II.

We extracted three-dimensional coordinates in ATOM lines of the PDB entry for each protein, calculated the distance d_{ij} between C_α atoms using the Euclidean distance, and transformed the distance matrix into the grayscale image each pixel of which has $\lfloor c \cdot d_{ij} \rfloor$ for a positive constant c because the value of a pixel must be an integer.

The source code was implemented by C++ using OpenCV library (version 2.4.9) [32]. All experiments were done in a single processing core of Xeon E5-2640 2.5GHz under linux operating system.

TABLE II. THE SIERK-PEARSON DATASET. IT CONTAINS 20 MAINLY ALPHA PROTEINS, 26 MAINLY BETA PROTEINS, AND 40 ALPHA-BETA PROTEINS.

Class of CATH	Proteins
Mainly alpha	1ad6A, 1ao6A5, 1bbhA0, 1cnsA1, 1d2zD0, 1dat00, 1e12A0, 1eqzE0, 1gwxA0, 1hgu00, 1hlm00, 1jnk02, 1mmoD0, 1nubA0, 1quuA1, 1repC1, 1sw6A0, 1trrA0, 2hpdA0, 2mtaC0
Mainly beta	1a8d02, 1a8h02, 1aozA3, 1b8mB0, 1bf203, 1bjqB0, 1bqyA2, 1btkB0, 1c1zA5, 1cl7H0, 1d3sA0,1danU0, 1dsyA0, 1dxmA0, 1et6A2, 1extB1, 1nfiC1, 1nukA0, 1otcA1, 1qdmA2, 1qe6D0, 1qfkL2, 1que01, 1rmg00, 1tmo04, 2tbvC0
Alpha-beta	1a1mA1, 1a2vA2, 1akn00, 1aqzB0, 1asyA2, 1atiA2, 1auq00, 1ax4A1, 1b0pA6, 1b2rA2, 1bcg00, 1bcmA1, 1bf5A4, 1bkcE0, 1bp7A0, 1c4kA2, 1cd2A0, 1cdg01, 1d0nA4, 1d4oA0, 1d7oA0, 1doi00, 1dy0A0, 1e2kB0, 1eccA1, 1fbnA0, 1gsoA3, 1mpyA2, 1obr00, 1p3801, 1pty00, 1qb7A0, 1qmvA0, 1urnA0, 1zfjA0, 2acy00, 2drpA1, 2nmtA2, 2reb01, 4mdhA2

TABLE III. RESULTS ON AUC BY THE SIMILARITY BASED ON SIFT AND SURF USING $c = 5, 10, 15$ FOR THE CHEW-KEDEM AND SIERK-PEARSON DATASETS, RESPECTIVELY.

Method		Chew-Kedem	Sierk-Pearson
SIFT-based	c = 5	0.99	0.60
	10	**1.0**	0.61
	15	**1.0**	0.61
SURF-based	c = 5	0.97	**0.63**
	10	0.96	0.62
	15	0.99	0.60
MSPC-ACD [14]		0.86	0.62
MSPC-NCD [14]		0.87	0.61
Krasnogor and Pelta [11]		0.80	0.50
Dai and Wang [31]		0.86	0.58

B. Results

We performed Receiver Operating Characteristic (ROC) analysis [33] as in [14], [31]. For evaluating the similarity based on SIFT and SURF, we considered the binary classification problem whether or not two protein structures are included in the same CATH class. In total, $\binom{n}{2}$ pairs were classified for a dataset of n proteins. Then, an ROC curve was plotted using the true positive rate, TP/(TP+FN), and the false positive rate, FP/(TN+FP), for a binary classifier as the threshold changes, where TP, FP, TN, and FN denote the numbers of true positives, false positives, true negatives, and false negatives, respectively, and the Area Under the Curve (AUC) was calculated.

Table III shows the results on AUC by the similarity based on SIFT and SURF using $c = 5, 10, 15$ for the Chew-Kedem and Sierk-Pearson datasets, respectively, where MSPC-ACD and MSPC-NCD mean the similarity measures of ACD and NCD using the MSPC image compression algorithm [14], respectively. We can see that the AUC value by the similarity based on SURF using $c = 5$ was larger than those by other existing similarity measures by Krasnogor and Pelta [11] and Dai and Wang [31], and those by MSPC-ACD and MSPC-NCD. For the Chew-Kedem dataset, the AUC values by the similarity based on SIFT using $c = 10, 15$ were largest.

Figures 2 and 3 show the results on ROC curves by the similarity based on SIFT and SURF using $c = 5, 10, 15$, ACD and NCD using MSPC for the Chew-Kedem and Sierk-Pearson datasets, respectively. For the Chew-Kedem dataset, our proposed similarity worked well. For the Sierk-Pearson

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

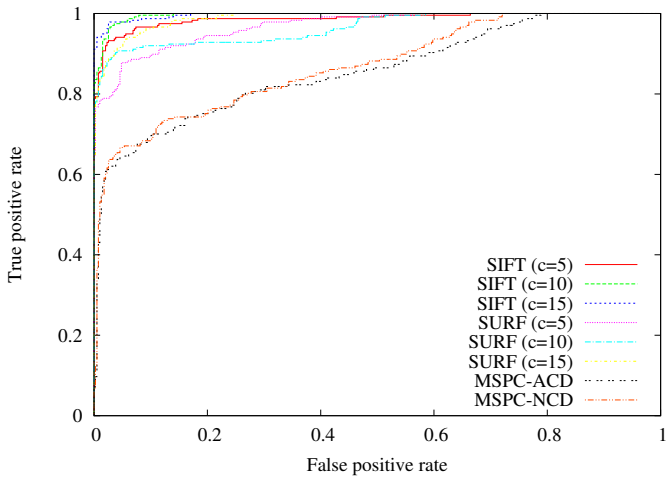


Fig. 2. Results on ROC curves by the similarity based on SIFT and SURF using $c = 5, 10, 15$, ACD and NCD using MSPC, respectively, for the Chew-Kedem dataset.

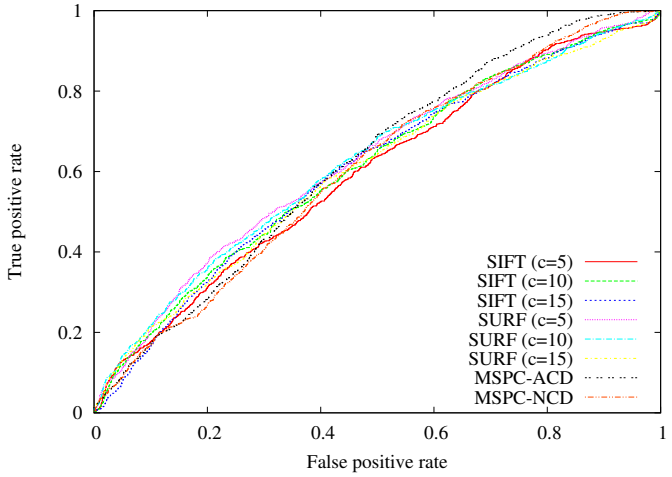


Fig. 3. Results on ROC curves by the similarity based on SIFT and SURF using $c = 5, 10, 15$, ACD and NCD using MSPC, respectively, for the Sierk-Pearson dataset.

dataset, the ROC curves were comparable with each other. However, we can see that the ROC curves by the similarity based on SURF using $c = 5, 10$ were better than others in the low range of false positive rates.

Figure 4 shows the result of single-linkage clustering using the similarity based on SIFT with $c = 15$ for the Chew-Kedem dataset, where the identifier in each parenthesis denotes the corresponding superfamily as shown in Table I. We can see that the proteins were well classified except the pair of 4enl01 (3.30.390.10) and 6xia (3.20.20.150).

Figure 5 shows the result on distribution of execution time (millisecond) of SIFT and SURF for two images using $c = 5, 10, 15$ over the Chew-Kedem and Sierk-Pearson datasets. In almost all cases, the execution time of SURF was shorter than that of SIFT, and both were less than 40 milliseconds.

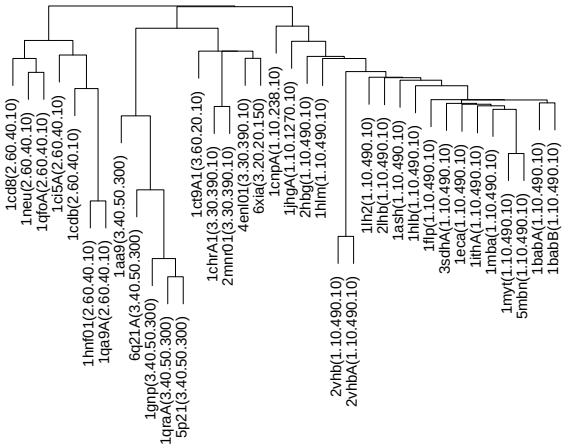


Fig. 4. Result of single-linkage clustering using the similarity based on SIFT with $c = 15$ for the Chew-Kedem dataset. The identifier in each parenthesis denotes the corresponding superfamily.

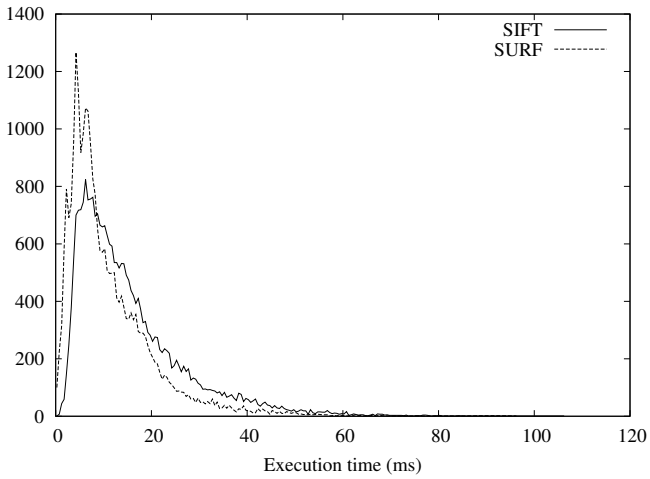


Fig. 5. Results on distribution of execution time (millisecond) of SIFT and SURF for two images using $c = 5, 10, 15$ over the Chew-Kedem and Sierk-Pearson datasets.

IV. Conclusion

We proposed the similarity between protein structures on the basis of the efficient and effective image local feature descriptors SIFT and SURF for image recognition, and applied it to two datasets of protein structures. The AUC value by the similarity based on SURF using $c = 5$ was larger than those by several existing similarity measures and by the compression-based similarity measures in our previous study, and the results suggest that SIFT and SURF are also useful for recognizing protein structures. The classification accuracy by our proposed similarity was very high for the Chew-Kedem dataset, whereas it was not so high for the Sierk-Pearson dataset. There, however, is much room to improve our proposed similarity

because we used SIFT and SURF without any modification and our similarity was defined in a simple manner. For instance, although the value of a pixel in images must be an integer, for our purpose, we can modify SIFT and SURF to allow values other than integers as their inputs. Other future work is to evaluate our methods using a more general dataset with more candidates, and to compare classification performance with other existing efficient similarity measures such as *i*SARST.

ACKNOWLEDGMENT

This work was partially supported by Grants-in-Aid #26240034, #24500361, and #26610037 from MEXT, Japan.

REFERENCES

[1] F. M. Pearl, C. F. Bennett, J. E. Bray, A. P. Harrison, N. Martin, A. Shepherd, I. Sillitoe, J. Thornton, and C. A. Orengo, "The CATH database: an extended protein family resource for structural and functional genomics," *Nucleic Acids Research*, vol. 31, pp. 452–455, 2003.

[2] A. Andreeva, D. Howorth, S. E. Brenner, T. J. P. Hubbard, C. Chothia, and A. G. Murzin, "SCOP database in 2004: refinements integrate structure and sequence family data," *Nucleic Acids Research*, vol. 32, pp. D226–D229, 2004.

[3] L. Holm and C. Sander, "Mapping the protein universe," *Science*, vol. 273, pp. 595–602, 1996.

[4] T. Akutsu, "Protein structure alignment using dynamic programming and iterative improvement," *IEICE Transactions on Information and Systems*, vol. E79-D, pp. 1629–1636, 1996.

[5] I. N. Shindyalov and P. E. Bourne, "Protein structure alignment by incremental combinatorial extension (CE) of the optimal path," *Protein Engineering*, vol. 11, pp. 739–747, 1998.

[6] W. Taylor, "Protein structure comparison using iterated double dynamic programming," *Protein Science*, vol. 8, no. 3, pp. 654–665, 1999.

[7] E. Krissinel and K. Henrick, "Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions," *Acta Crystallogr D*, vol. 60, pp. 2256–2268, 2004.

[8] Y. Zhang and J. Skolnick, "TM-align: a protein structure alignment algorithm based on the TM-score," *Nucleic Acids Research*, vol. 33, pp. 2302–2309, 2005.

[9] S. Minami, K. Sawada, and G. Chikenji, "MICAN: a protein structure alignment algorithm that can handle Multiple-chains, Inverse alignments, C_α only models, Alternative alignments, and Non-sequential alignments," *BMC Bioinformatics*, vol. 14, p. 24, 2013.

[10] K. Henrick, Z. Feng, W. F. Bluhm, D. Dimitropoulos, J. F. Doreleijers, S. Dutta, J. L. Flippen-Anderson, J. Ionides, C. Kamada, E. Krissinel, C. L. Lawson, J. L. Markley, H. Nakamura, R. Newman, Y. Shimizu, J. Swaminathan, S. Velankar, J. Ory, E. L. Ulrich, W. Vranken, J. Westbrook, R. Yamashita, H. Yang, J. Young, M. Yousufuddin, and H. M. Berman, "Remediation of the protein data bank archive," *Nucleic Acids Research*, vol. 36, pp. D426–433, 2008.

[11] N. Krasnogor and D. A. Pelta, "Measuring the similarity of protein structures by means of the universal similarity metric," *Bioinformatics*, vol. 20, pp. 1015–1021, 2004.

[12] A. Caprara, R. Carr, S. Istrail, G. Lancia, and B. Walenz, "1001 Optimal PDB structure alignments: Integer programming methods for finding the maximum contact map overlap," *Journal of Computational Biology*, vol. 11, pp. 27–52, 2004.

[13] M. Li, J. H. Badger, X. Chen, S. Kwong, P. E. Kearney, and H. Zhang, "An information-based sequence distance and its application to whole mitochondrial genome phylogeny," *Bioinformatics*, vol. 17, pp. 149–154, 2001.

[14] M. Hayashida and T. Akutsu, "Measuring the similarity of protein structures using image compression algorithms," *IEICE Transactions on Information and Systems*, vol. E94-D, no. 12, pp. 2468–2478, 2011.

[15] M. Polvere and M. Nappi, "A feature vector technique for fast fractal image coding," University of Salerno, Tech. Rep., 1998.

[16] A. Said and W. A. Pearlman, "Reversible image compression via multiresolution representation and predictive coding," in *Proc. SPIE*, vol. 2094, 1993, pp. 664–674.

[17] J. Rocha, F. Rosselló, and J. Segura, "The Universal Similarity Metric does not detect domain similarity," 2008. [Online]. Available: http://arxiv.org/pdf/q-bio/0603007v1.pdf

[18] O. Carugo and S. Pongor, "Protein fold similarity estimated by a probabilistic approach based on C(alpha)-C(alpha) distance comparison," *Journal of Molecular Biology*, vol. 315, no. 4, pp. 887–898, 2002.

[19] P. Røgen and B. Fain, "Automatic classification of protein structure by using Gauss integrals," *Proc. Natl. Acad. Sci. USA*, vol. 100, no. 1, pp. 119–124, 2003.

[20] I. Choi, J. Kwon, and S. Kim, "Local feature frequency profile: A method to measure structural similarity in proteins," *Proc. Natl. Acad. Sci. USA*, vol. 101, no. 11, pp. 3797–3802, 2004.

[21] E. Zotenko, D. O'Leary, and T. Przytycka, "Secondary structure spatial conformation footprint: A novel method for fast protein structure comparison and classification," *BMC Structural Biology*, vol. 6, p. 12, 2006.

[22] W.-C. Lo, C.-Y. Lee, C.-C. Lee, and P.-C. Lyu, "iSARST: an integrated SARST web server for rapid protein structural similarity searches," *Nucleic Acids Research*, vol. 37, pp. W545–W551, 2009.

[23] W.-C. Lo, P.-J. Huang, C.-H. Chang, and P.-C. Lyu, "Protein structural similarity search by Ramachandran codes," *BMC Bioinformatics*, vol. 8, p. 307, 2007.

[24] I. Budowski-Tal, Y. Nov, and R. Kolodny, "FragBag, an accurate representation of protein structure, retrieves structural neighbors from the entire pdb quickly and accurately," *Proc. Natl. Acad. Sci. USA*, vol. 107, no. 8, pp. 3481–3486, 2010.

[25] K. Molloy, M. Van, D. Barbara, and A. Shehu, "Exploring representations of protein structure for automated remote homology detection and mapping of protein structure space," *BMC Bioinformatics*, vol. 15, no. Suppl 8, p. S4, 2014.

[26] D. Lowe, "Object recognition from local scale-invariant features," in *IEEE International Conference on Computer Vision*, 1999, pp. 1150–1157.

[27] ——, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[28] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-Up Robust Features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[29] L. P. Chew and K. Kedem, "Finding consensus shape for a protein family," *Algorithmica*, vol. 38, pp. 115–129, 2003.

[30] M. L. Sierk and W. R. Pearson, "Sensitivity and selectivity in protein structure comparison," *Protein Science*, vol. 13, pp. 773–785, 2004.

[31] Q. Dai and T. Wang, "Comparison study on k-word statistical measures for protein: From sequence to 'sequence space'," *BMC Bioinformatics*, vol. 9, p. 394, 2008.

[32] http://opencv.org/.

[33] C. E. Metz, "ROC methodology in radiologic imaging," *Investigative Radiology*, vol. 21, pp. 720–733, 1986.

978-1-4799-7294-4/14/$31.00 ©2014 IEEE