International Journal of Engineering &Technology, 7 (1) (2018) 110-114

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i1.9009
Research paper

Applying metric space and pivot-based indexing on combined features of bio-images for fast execution of composite queries

Meenakshi Srivastava [1]*, S.K. Singh[2], S.Q Abbas [3]

[1]Amity Institute of Information Technology, Amity University, Uttar Pradesh, India
[2]Amity Institute of Information Technology, Amity University, Uttar Pradesh, India
[3]Computer Science Department, Ambalika Institute of Information Technology, Uttar Pradesh, India
**Corresponding author E-mail: msrivastava@lko.amity.edu*

Abstract

Content based recovery of bio images requires index structures, which can retrieve the similar image objects in time proficient way. Conventional Structure/ sequence based recovery of bio-images (for example, protein structures) experiences, tedious online similarity check from huge web based databases. The general approach of image feature representations follows vector based portrayal. In present manuscript, visual highlights of 3D protein structures and their content highlights have been implemented in isolated metric space, rather than vector space which advances the similarity recovery. At long last, the Visual highlights and Content based highlights are consolidated in one metric space, through the component results of highlight and substance metric. Results have demonstrated that pivot based ordering/ indexing on Combined Index Metric can undoubtedly execute composite content construct queries with respect to bio images in time effective way.

Keywords: *Image Retrieval; Metric Space; Protein Structures; AESA; LAESA; Pivot Indexing.*

1. Introduction

Content based multimedia retrieval methods focus on similarity search rather than exact search. In Content Based Image Retrieval, there are two methodologies for recovery from the database, the primary approach manages recovery through explanations that is metadata, which depicts the picture and the second approach utilizes the picture question itself for the seeking reason [1]. The keyword based retrieval faces semantic gap, and in many cases fails to retrieve the relevant information. The second approach of image retrieval stores the visual information of image like color, shape or texture in the database using the feature vectors [2]. Various researchers [3-4, 6-8] used vector space's representation of geometric properties of multimedia object. Searching is done by matching the feature vector of the query image with feature vectors of the image objects stored within the database. Matching determines the similarity, which is done by computing a distance on the feature vectors. Exact match retrieval, is not enough or practical for the areas like image databases, text documents, audio and video collections, or bio images databank, etc. For the better results searching should be based on standard form of closeness, similarity, or dissimilarity between query and the objects in the database. In response to a query, a query response set is formed, this set contains objects that are close to the given query object. – indexing plays an important role in searching algorithm, as they build a data structure to speed up the search. Indexing algorithms performs well in low-dimensional spaces, as higher dimensions based index structures [9-11] on average stop being efficient, when the dimensionality exceeds to twenty. The quality of searching algorithm can be measured on many criteria like:(a) total number of distance computation required during a query, (b) number of required disk accesses and the CPU time used further than (a) or (b). Distance Computations are very costly in the case o complex objects such as 3D images, as the number of computations directly affects the run time cost. An alternative method of modeling such as complex multimedia data is through metric space rather than the vector space so that the run time similarity distance computation is fast. In metric space, the similarity is computed by a positive distance function which provides a concept for the nearness. In the present manuscript similarity distance between the images of proteins, structures are represented via Euclidian Distance based metric space. Metric space based storage facilitates fast searching through the database via pivot based indexing algorithm.

The manuscript is organized as follows. Problem Statement and State of Art has been discussed in second Section 2. Section 3 covers pivot based indexing and searching. Section 5 contains Results and discussion of AESA performance. Finally, the conclusion and future work appear in Section 5.

2. State of art

Advances in research focus methodologies to choose the structures of bio-particles have prompted a huge increase in the sizes of the protein structure databases, for instance, Protein Data Bank (PDB) [12]. In 1992, only 1,000 structures were stored in PDB, whereas in 2002 the number of structures was over 18,000 and in 2017, there are more than 103,514 structures in the PDB. The existing methods of similar protein retrieval from the structural databases are penalized due to lack of fast searching algorithms. Most of the existing methods are based on structural alignment, which won't be a preferred choice for protein structure search against the large database, since it is computationally expensive to compute their similarities [13]. Tools and web servers such as Clustal series, T

coffee, BLAST (Basic Local Alignment Search Tool), FASTA, HMMER, etc. are good at the sequence alignment whereas tools such as MAMMOTH, Dali Lite, CE(Combinatorial extension) etc. are used frequently by the scientist for structural alignment. Although some of these tools are linked and on the request from the user, the data gets transferred from one tool/site to the other for further analysis, but this approach needs improvement for a better and a faster analysis of the structural and the sequential information of proteins at hand. Understanding protein similarity relationships is vital for the Annotation of genome sequences (Andrade et al., 1999; Pearl et al., 2000; Wilson et al., 2000; Todd et al., 2001). Proteins having high sequence identity and high structural similarity tend to possess functional similarity and evolutionary relationships, yet examples of proteins deviating from this general relationship of sequence/structure/ function homology are well-recognized. Varied sequence/structure similarity relationships werereported by various researchers. For example, high sequence identity but low structure similarity can occur due to conformational plasticity, mutations, solvent effects, and ligand binding, etc. Most of the present work has focused on the expected similarity relationship where the proteins have significant sequence and structural similarity. Wilson et al., 2000; Chothia and Lesk, 1986; Russell et al., 1997; Levitt and Gerstein, 1998; Wood and Pearson, 1999). Extra effort and funds are currently being invested to improve and speed-up the processing potential of many computer-based tools that reign in the field of structural bioinformatics [5]. In [14], a novel method for extraction of visual features from the PDB files using the intelligent vision algorithm has been implemented. In [15] content based server 'AMIPRO' has been implemented using intelligent vision algorithm proposed in [14]. In AMIPRO [15] High Order Autocorrelation (HLAC) features had been used for extraction of visual features from 3 D protein images, and protein sequence alignment algorithm was used for calculating content similarity. The present manuscript extends the work done in [15] by applying pivot based indexing on metric of combined features. The proposed Combined Index Metric based indexing can easily retrieve structure and sequence based similar proteins in time efficient manner. A brief description of visual feature extraction of AMI-PRO [15] has been done in 2.1, and in 2.2 basic property of metric space is detailed.

2.1. Visual feature extraction

The size of protein image has been fixed to 128 x 128 pixels using JMOL software [19]. For geometrical feature extraction an intelligent vision algorithm proposed in [16-17] has been deployed. Geometrical Feature Extraction concerns the extraction of features which are invariant under some transformation group acting on pattern [17]. The primitive features for an intelligent vision must be shift Invariant and Additive. The autocorrelation function can easily extract Shift Invariant and additive featured. For extracting function High Order Local Autocorrelation (HLAC) function is used.

Each supplied query image is rotated randomly around its three principal viewing axes and multiple-views of 2D images are stored. 2D HLAC (High-Order Order Local Autocorrelation) features [17] are extracted from the query images. Duplicate configurations are removed, and local mask patterns are reduced to 35. The combined HLAC features produce a 105-dimensional HLAC feature vector. Principal Component Analysis is performed on HLAC feature vector and Eigen value of the covariance of the matrix x [M N] has been calculated. Next the Euclidian Distance between two Eigen vectors is computed. This distance represents the similarity between two protein structures.

2.2. Metric space property

A metric space is a set P defined as a function d: P ×P → R which measures the distance d (p, q) between points p, q ∈ P. A function f satisfying the Positivity, Symmetry and Triangle Equality property on P is called a metric on P. Positivity, Symmetry and Triangle Equality of a metric pace P can be summarized as-

i). Positivity is defined as, that for all p, q ∈ P, d (p, r) ≥ 0 with equality if and only if p = r.
ii). Symmetry defines that for all p, q ∈ P, d(p, q) = d (q, p)
iii). Triangle Equality states that for all p, q, r ∈ P d(p, q) ≤d (p, r) + d (r, q)

3. Pivot based indexing and searching

Vidal, 1986 introduced Approximating and Eliminating Search Algorithm AESA, which is a pivot-based metric space search algorithm. For two decades (Figueroa et al., 2009), AESA is being considered the fastest NN search methods in metric spaces [16]. The pivots are a subset of objects of the database that are used to speed up the search. Nearest Neighbor (NN) search are based on similarity search, and the measured dissimilarity is interpreted as a distance.

In order to find the Nearest Neighbor AESA applies two iterating steps: at first step a candidate to NN is selected and at second step, the selected candidate is used to discard all those database's objects which have the greater distance value than the current candidate. Performance of AESA degrades when the data set is large; to overcome this, we have divided the data set into clusters and selection of the appropriate cluster for searching is the first step of our implementation AESA. LAESA [15] the Linear Approximating and Eliminating Search Algorithm was introduced to overcome the data set quadratic size constraint of AESA, but LAESA suffers with additional preprocessing time and linear growth in memory size with the prototype.

3.1. Cluster based implementation of AESA

AESA selects the basic similarity value randomly to and then starts the search for Nearest Neighbor based on computation of lower bound, whereas in our approach at first step the, the distance between the query point and the center of each cluster d (q, c) is calculated and the cluster which have the minimum distance from the query point is selected as base property for NN search.

3.2. Algorithm basic similarity BS - selection

Deriving Linear searching strategy is possible using Branch and Bound algorithms. The basic difference is the bounding function reliability based on feature vector like Euclidian distance as elements of the database in the form of two-dimensional arrays. The basic square matrices (n x n) are obtained based on different feature vector's spaces like Euclidian distance and text similarity and so on. Let S be the set of similarity values and B⊆ S the set of Basic Similarity values. Let x be a test image and Q ⊆S be a set of similarity values q for which d(x, q) can be com put (and stored) of the search procedure. Then for every s € S, we can apply the selection of basic similarity asset value's algorithm as follows.

Entry elements: S⊂ E; m € N; {a set (finite) of similarity values and number of Basic similarities}

Result obtained: B⊂S, $|B|$=m{a set of m Base Similarity values (BSs)}

ED€ $R^{|S| \times |B|}$; { $|S| - |B|$ inter similarity values Euclidian distances}

Functions: ed: E X E→ R; {Euclidian distance function}

Key role players/Variables:A € $R^{|S|}$;{Euclidian distance array of accumulator}

b, b '€S; max€ IR;

begin

b' := any arbitrary image element (S); B:={b'}; A:= [0];

While | B | < m do

{

max: = 0; b:=b'
For every s € S – B do
ED [b, s]: = d(b, s);
A[s]: = A[s] +ED [b, s];
If {A[S] >max) then
b' :=s;
max :=A [s];
}
En dif
End of for loop
B: =B ∪{b'};
}
End of while loop
End

The computational complexity of this algorithm is n .m steps (each involving one Euclidian distance computation and other elementary unit-cost operations), where n= | S | is the number of similarity values and m = | B | is the given number of Base Similarity values.

4. Results and discussions

To check the performance metric space model on the real data set [14] collected from RCSB PDB a series of experiments was carried out. Our data set is classified into four classes of SCOP database, i.e. Alpha (α), Beta (β), Alpha/Beta (α/β), and Alpha + Beta (α/β), so to reduce the search time, instead of searching in the whole database the distance of query image with the cluster centers of each class is measured and the query object is searched into the clusters for which the calculated distance was measured to be minimum. The object which has the maximum distance with the cluster center has been chosen as the candidate for Base Similarity. Since our main aim is to perform content based retrieval two AESA metric structure, one for Visual similarity distance(ED) and second for Content based similarities (CD) are created. A combined Index structure is also generated by performing element based multiplication of ED and CD.

4.1. Performance analysis of AESA

1) AESA [16] stores a metric of distances between database objects. Distance between the all object is computed at the time of creation of AESA. The structure of the ASEA matrix is n × n, but half of the matrix below the diagonal is stored. That is, n (n − 1)/2 distances, because the computed distance matrix satisfies the matrix property and the elements above, and below the diagonal are same.
2) For search operation for range query R (q, r) our implementation of AESA picks up an object, for example, I1 (Base Similarity) which has the maximum distance from the cluster center. The exact Euclidian distance between I1 and Q1 is computed let's say O. Now this distance will be used for pruning objects.
3) Pruning of object I if |d(I, O) − d(q, p)| > r, that is, the lower bound in is greater than the query
4) The next pivot is chosen among all non discarded objects up to now.
5) The process is stopped when the set of non-discarded objects is small enough.
6) Lastly, the distance of remaining objects are directly compared with q, and objects with d (q, o) ≤ r are reported.

4.2. Generation of combined index metric space

The metric ED represents the Euclidian Distance based visual similarity between all the object of the database. In our case the content similarity refers to the sequence based similarity between any two proteins, which is always represented in the form of percentage like 50%, 80% etc. A metric CD, having the distance between the sequence similarities on same protein objects as in ED is also created. To normalize the CD metric the percentage value has been represented on the scale of 1 i.e. 50% similarity will be stored - as 0 .5 and 80 % similarity will be stored as 0.8. Now to perform the content based query like "Which are the proteins that are 50% structurally similar and 70% sequentially similar", combined search on both of distance matrices is required. Since the matrices are square matrices element product metrics generation is possible. This way of indexing, in turn will reduce the time taken separately on two individual element metric indexing. The Oder of n X n square metrics in each case of CD and ED will be in the range of O $(n)^2$ i.e. total 2 x O $(n)^2$ whereas in the product metrics the order will remain n^2 thus reducing the time by ½. In general for N feature similarity checking the time will reduce to the extent of 1/N.

	I_1	T_2	T_3	I_4	I_5
I_1	0	2	1	5	8
I_2	2	0	3	7	5
I_3	1	3	0	4	6
I_4	5	7	4	0	11
I_5	8	5	6	11	0

Fig.1: Euclidian Distance Combined Index Matrix.

The results show that raising the number of dimensions, does not affect the average dimensions number of distance metric evaluations done by ASEA. Our cluster based implementation had provided a good Base similarity value as the selected candidate had the maximum distance from the rest of the objects.
It's very obvious from the performance bar diagram of AESA and LAESA shown in Fig. 5, that the no. of distance studied for various dimensions is always lower for the lowest one of the LAESA. It can be inferred that the distance in case of LAESA, shows a considerably higher degree of dependence on the number of dimensions. Although, the trend observed in LAESA also depicts an increase in distance with increase in the number of dimensions, the amount of increase is massive for LAESA as compared to AESA.

5. Conclusion and future work

In the present manuscript metric space based representation of visual features and content based features of bio images has been discussed in context of 3D protein structures. Performance of cluster oriented AESA and LAESA on protein image has been measured. In Comparison of AESA and LAESA, performance of AESA was better than LAESA. Metric space based representation of data involves pre computation of distance between the object, and AESA used pivot based method for searching similar object. Though AESA algorithm suffers with quadratic space complexity O (n^2) and quadratic construction complexity, then also in the tradeoff between space and time, we prefer fast searching because one online distance computation is much more expensive than one scan in metric. Secondly the cluster based implementation has reduced the quadratic effect to an extent.

Table 1: Average Number of Distances Computed by AESA, LAESA (Using 8, 12 and 16 Pivots for Dimensions 4, 6 and 8 Respectively)

Dimension (size of data set/ Method	Dimension 4			Dimension 6			Dimension 8		
	100	150	200	100	150	200	100	150	200
AESA	3.45	3.30	3.64	3.45	3.31	3.65	3.45	3.32	3.66
LAESA	4.42	4.02	3.89	5.23	5.66	5.45	6.88	7.02	7.89

Table 2: Average Number of Distances Computed by AESA and LAESA Algorithms Using a Training Set of 200 Objects and 72 Queries with Databases

Size of Database Method	100	150	200
AESA	3.45	3.47	3.48
LAESA	5.51	5.67	5.74

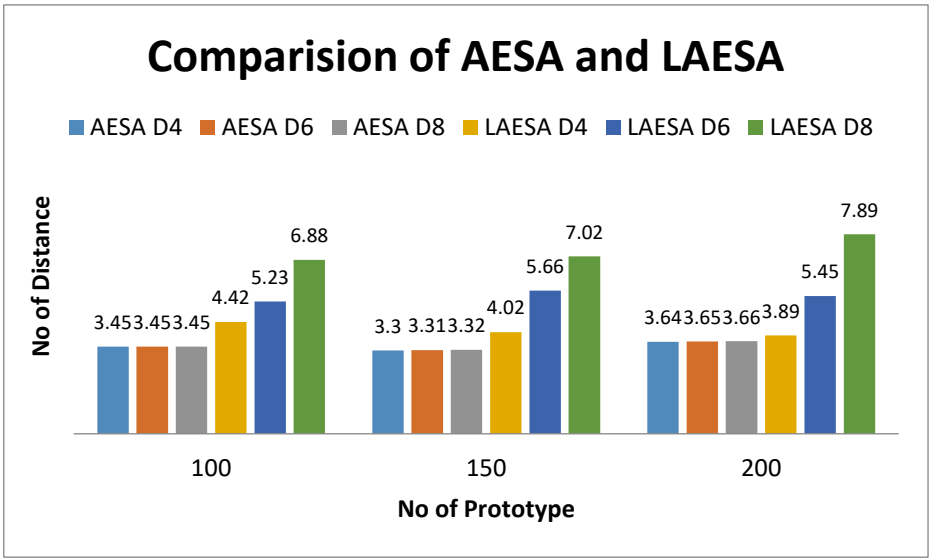

Fig. 5: Average Number of Distance Computations in LAESA as A Function of the Number of Prototypes for D = 4 and D=6 and D = 8.

The Combined Index Metric space which is created via element based product metric of feature metric and content metric can retrieve the result easily for the queries which involve feature and content based combined searching. Our future work involves development of better cluster based implementation of AESA so that quadratic space complexity can be minimized without compromising the retrieval speed. We will implement the proposed combined index metric in other image retrieval fields of science and research disciplines including Earth science, materials science, biology, and medicine.

References

[1] Mussarat Yasmin, Sajjad Mohsin, Muhammad Sharif, Intelligent Image Retrieval Techniques: A Survey ,Journal of Applied research and technology, Vol 14, 2016

[2] JohanW. H. Tangelder · Remco C. Veltkamp, "A survey of content based 3D shape retrieval methods", Multimed Tools Appl (2008)

[3] Paolo Ciaccia, Marco Patella, and Pavel Zezula. A cost model for similarity queries in metric spaces. In Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 1-3, 1998, Seattle, Washington, pages 59–68. ACM Press, 1998.https://doi.org/10.1145/275487.275495.

[4] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. Communications of the ACM, 18(9):509–517, 1975.https://doi.org/10.1145/361002.361007.

[5] https://www-roc.inria.fr/gamma/gamma/Logiciels/index.en.html

[6] Jon Louis Bentley. Multidimensional binary search trees in database applications. IEEE Transactions on Software Engineering, 5(4):333–340, 1979.https://doi.org/10.1109/TSE.1979.234200.

[7] Antonin Guttman. R-Trees: A dynamic index structure for spatial searching. In Beatrice Yormark, editor, SIGMOD'84, Proceedings of Annual Meeting, Boston, Massachusetts, pages 47–57. ACM Press, 1984.https://doi.org/10.1145/602259.602266.

[8] Hanan Samet. The quadtree and related hierarchical data structures. ACM Computing Surveys (CSUR), 16(2):187–260, 1984.https://doi.org/10.1145/356924.356930.

[9] Stefan Berchtold, Daniel A. Keim, and Hans-Peter Kriegel. The X-tree: An index structure for high-dimensional data. In T. M. Vijayaraman, Alejandro P. Buchmann, C. Mohan, and Nandlal L. Sarda, editors, VLDB'96, Proceedings of 22th International Conference on Very Large Data Bases, September 3-6, 1996, Mumbai (Bombay), India, pages 28–39. Morgan Kaufmann, 1996.

[10] Andreas Henrich. The LSDh-Tree: An access structure for feature vectors. In Proceedings of the Fourteenth International Conference on Data Engineering, February 23-27, 1998, Orlando, Florida, USA, pages 362–369. IEEE Computer Society, 1998.https://doi.org/10.1109/ICDE.1998.655799.

[11] Kaushik Chakrabarti and Sharad Mehrotra. The Hybrid Tree: An index structure for high dimensional feature spaces. In Proceedings of the 15th International Conference on Data Engineering, 23-26 March 1999, Sydney, Austrialia, pages 440–447. IEEE Computer Society, 1999.

[12] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E.Bourne. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000.https://doi.org/10.1093/nar/28.1.235.

[13] Sungchul Kim, Postech, York Korea, Fast Protein 3D Surface Search, *ICUIMC (IMCOM)' 13,* January 17-19, 2013, Kota Kinabalu, Malaysia Copyright 2013 ACM 978-1-4503-1958-4.

[14] Meenakshi Srivastava, Dr. S.K.Singh, Dr. S.Q.Abbas, "A Novel Model for Fast And Robust Retrieval of 3D Bio-Images Using Intelligent Vision Algorithm", International Journal of Control Theory and Applications, 9(41) 2016, pp. 617-627.

[15] Srivastava M., Singh S.K., Abbas S.Q., Neelabh (2018) *AMIPRO*: A Content-Based Search Engine for Fast and Efficient Retrieval of 3D Protein Structures., Smart Innovation, Systems and Technologies, Springer, Volume 79,https://doi.org/10.1007/978-981-10-5828-8_70.

[16] Marla Luisa Mic6, Jos6 Oncina, A new version of the Nearest - Neighbour Approximating and Eliminating Search Algorithm (AESA) with linear preprocessing time and memory requirements, Pattern Recognition Letter, 1994, revised 2011.

[17] Motofumi T. Suzuki, Texture Image Classification using Extended 2D HLAC Features,KEER2014, LINKÖPING | JUNE 11-132014 International Conference On Kansai Engineering And Emotion Research Texture
[18] Nobuyuki OTSU and Takio Kurita, A New Scheme for Practical Flexible And Intelligent Vision Systems, lAPR Workshop on CV - Special Hardware and Industrial Applications OCT.12-14, 1988, Tokyo.
[19] A. Herr´aez. Biomolecules in the computer: Jmol to the rescue. *Biochemistry and Molecular Biology Education*, 34(4):255–261, 2006.https://doi.org/10.1002/bmb.2006.494034042644.