Mu et al. BMC Bioinformatics (2019) 20:351
https://doi.org/10.1186/s12859-019-2943-x

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

DCGR: feature extractions from protein sequences based on CGR via remodeling multiple information

Zengchao Mu[1†], Ting Yu[1†], Enfeng Qi[2], Juntao Liu[1*] and Guojun Li[1*]

Abstract

Background: Protein feature extraction plays an important role in the areas of similarity analysis of protein sequences and prediction of protein structures, functions and interactions. The feature extraction based on graphical representation is one of the most effective and efficient ways. However, most existing methods suffer limitations from their method design.

Results: We introduce DCGR, a novel method for extracting features from protein sequences based on the chaos game representation, which is developed by constructing CGR curves of protein sequences according to physicochemical properties of amino acids, followed by converting the CGR curves into multi-dimensional feature vectors by using the distributions of points in CGR images. Tested on five data sets, DCGR was significantly superior to the state-of-the-art feature extraction methods.

Conclusion: The DCGR is practically powerful for extracting effective features from protein sequences, and therefore important in similarity analysis of protein sequences, study of protein-protein interactions and prediction of protein functions. It is freely available at https://sourceforge.net/projects/transcriptomeassembly/files/Feature%20Extraction.

Keywords: Protein feature extraction, CGR curve, Physicochemical property, Algorithm

Background

Similarity analysis of protein sequences plays an important role in protein sequence studies, e.g. the prediction or classification of protein structures and functions. In General, the biological function of a protein is determined by its three dimensional structure which is dependent on the linear sequence of amino acids. Rigden [1] presented that one of the fundamental principles of molecular biology is that proteins having similar sequences possess similar functions. Up to now, lots of methods have been proposed for the similarity analysis of protein sequences, among which the graphical representation of protein sequences is one of the most used and effective strategies [2–21].

The chaos game representation (CGR) based on an iterative function system was firstly proposed for the representation of DNA sequences by Jeffrey in 1990 [22]. The Jeffrey's CGR is drawn within a quadrate with four vertices referring to nucleotides A, C, G and T. The first point is placed halfway between the center of the quadrate and the vertex corresponding to the first nucleotide of the sequence. The i-th ($i > 1$) point is placed halfway between the (i-1)-th point and the vertex corresponding to the i-th nucleotide. Being capable of discovering the inner pattern of gene sequences, CGR has been widely used in the investigation of DNA sequences [23–28]. Encouraged by the CGR of DNA sequences, the CGR of protein sequences has also been extensively studied by many researchers. Fisher et al. [29] first proposed an improved CGR of protein sequences, which was produced in a 20-side regular polygon with 20-vertices representing 20 kinds of amino acid. Randić et al. [30] constructed the CGR of protein sequences in the interior of a unit circle, on the circumference of which 20 amino acids are located uniformly according to the alphabet order of their three letter codes.

Amino acids themselves have physicochemical properties, which are important for protein structures, functions and

* Correspondence: juntaosdu@126.com; guojunsdu@gmail.com
†Zengchao Mu and Ting Yu contributed equally to this work.
[1]School of Mathematics, Shandong University, Jinan 250100, Shandong Province, China
Full list of author information is available at the end of the article

protein-protein interactions and have strong effects on the pattern of protein evolution. Therefore, physicochemical properties of amino acids have been widely used in protein sequence studies, such as similarity analysis of protein sequences, prediction of protein subcellular localization and protein structural class prediction [2–15, 18–20, 31–38]. In [39], Randić mentioned that ordering amino acids based on their physicochemical properties may offer better insights in comparative studies of proteins than representations of proteins based on alphabetical ordering of amino acids, which is essentially equivalent to random ordering. Following Randić's approach, He et al. [31, 40] proposed some different cyclic orders for the 20 amino acids to introduce the CGRs of protein sequences based on the physicochemical properties of amino acids. We denote the above CGRs by 20-CGR as 20 kinds of letters are used to represent protein sequences. Basu et al. [41] used a 12-sided regular polygon to generate the 12-CGR of protein sequences, each vertex of which represents a group of amino acids based on the conservative substitutions. Later Yu et al. [32] and Manikandakumar et al. [33] proposed 4-CGR, 5-CGR and 6-CGR for protein sequences, in which 4, 5 and 6 kinds of letters were used to represent protein sequences, respectively. In fact, using reduced amino acid alphabet to represent a protein sequence would easily result in loss of sequence information, since the amino acids belonging to the same group are considered identical.

So far, CGR method has achieved many applications in the studies of bioinformatics. The key issue in the application of CGR is to extract as many useful features as possible from CGR and several studies showed that those extracted features plays important roles in protein studies [25–28, 31, 34–38, 40–42]. One of the most frequently used feature extraction methods is the so-called FCGR, in which the CGR image is split into small grids and the frequencies of points falling into each grid are taken as the feature of the corresponding protein sequence. For example, in [34–38, 41], the CGR image of a protein sequence was split into 24 grids, and the frequencies of points falling into 24 grids are counted and taken as the numerical characteristics of the protein sequence. Under this procedure, a protein sequence can be converted into a 24-dimensional vector. Although FCGR method could effectively extract useful information from CGR, however, it loses the distribution information of the points in each grid, which is proved of great importance in this paper.

In this paper, we propose a novel feature extraction method of protein sequences based on the Randić's 20-CGR, which effectively integrates the physicochemical properties of amino acids into the construction of CGR curves and makes full use of the distribution information of points for extracting numerical characteristics from CGR curves. When tested on five data sets, it performs much better than all the compared methods.

Results

In this study, five most frequently used data sets were adopted to evaluate the performance of the new method DCGR in comparison with different feature extraction methods and also the sequence alignment method ClustalW.

Similarity analysis of 9 ND5 protein sequences

We first apply DCGR to analyze the similarities of the ND5 protein sequences from 9 kinds of species (detailed in Additional file 1: Table S1), which have been widely used in different studies and considered as a standard to evaluate the model [2, 4–8, 10, 12–15, 19, 20, 31, 43].

Based on DCGR, we first obtained a 9×632 feature matrix for the 9 protein sequences. Then PCA was used to reduce the dimensionality of the feature vectors. Here, only the first 6 principal components were selected and therefore a 9×6 reduced feature matrix could be built. *Euclidean* distance was used to calculate the distance between each two protein sequences (see Additional file 1: Table S2 for the calculated distances between protein sequences). The smaller distance between two proteins, the closer relationship between the two species.

From Additional file 1: Table S2, it is clear that the distance between Fin whale and Blue whale is the smallest of all, demonstrating the closest phylogenetic relationship between them. The distances among Human, Pigmy chimpanzee, Common chimpanzee and Gorilla are also small, showing that they are also similar. In addition, we can also find that Rat and Mouse have a relatively close relationship. However, the distance between Opossum and any other 8 species was very large, demonstrating its far relationship with the others. All results are consistent with the known evolutionary relationship among the 9 species.

For direct survey of evolutionary relationship among the 9 species, we construct the phylogenetic tree based on the distance matrix in Additional file 1: Table S2 shown in Fig. 1, which clearly illustrates four different branches clustered from the 9 species. The first branch consists of the Rodentia (Rat, Mouse), the second one the Primates (Pigmy chimpanzee, Common chimpanzee, Human, Gorilla), the third one the Cetacea (Fin whale and Blue whale) and the fourth one the Marsupialia (Opossum). ClustalW is one of the most popular multiple sequence alignment methods. Here, we also construct the phylogenetic tree by using ClustalW shown in Additional file 1: Figure S1, which shows very similar evolutionary relationships of the 9 species with our results.

Similarity analysis of 36 protein sequences

In the second example, we apply our method to analyze a data set consisting of 36 protein sequences of 5 different families: Globin (1eca, 5mbn, 1hlb, 1hlm, 1babA, 1babB, 1ithA, 1mba, 2hbg, 2lhb, 3sdhA, 1ash, 1flp, 1myt,

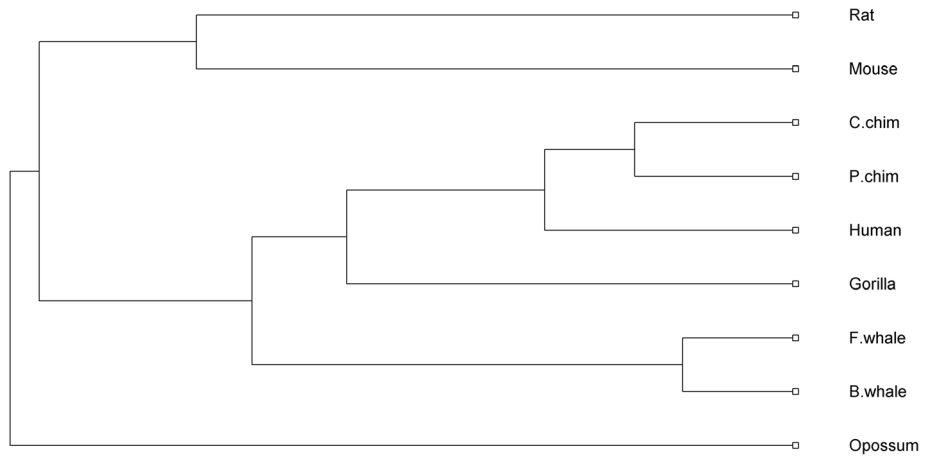

Fig. 1 Phylogenetic tree of the nine ND5 proteins constructed by DCGR

1lh2, 2vhbA, 2vhb), Alpha–Beta (1aa9, 1gnp, 6q21A, 1ct9A, 1qraA, 5p21), Tim-Barrel (6xia, 2mnr, 1chrA, 4enl), Beta (1 cd8, 1ci5, 1qa9, 1cdb, 1neu, 1qfoA, 1hnf), and Alpha (1cnp, 1jhg) [20, 43–48]. After extracting features by the method DCGR and reducing the dimensionality using PCA, the *Manhattan* distance was used to calculate the distance matrix of the 36 protein sequences. Similarly, we constructed the phylogenetic tree of the 36 protein sequences in Fig. 2, demonstrating that the 36 proteins have been accurately clustered into the 5 corresponding families, with only one erroneously clustered protein 1ct9.

In order to illustrate the superiority of DCGR, we compared its performance with six other methods including ClustalW in [20, 43–47], and the phylogenetic trees constructed by the six methods have been shown in Additional file 1: Figures S2-S8. After comparison, DCGR showed best performance since most of the six

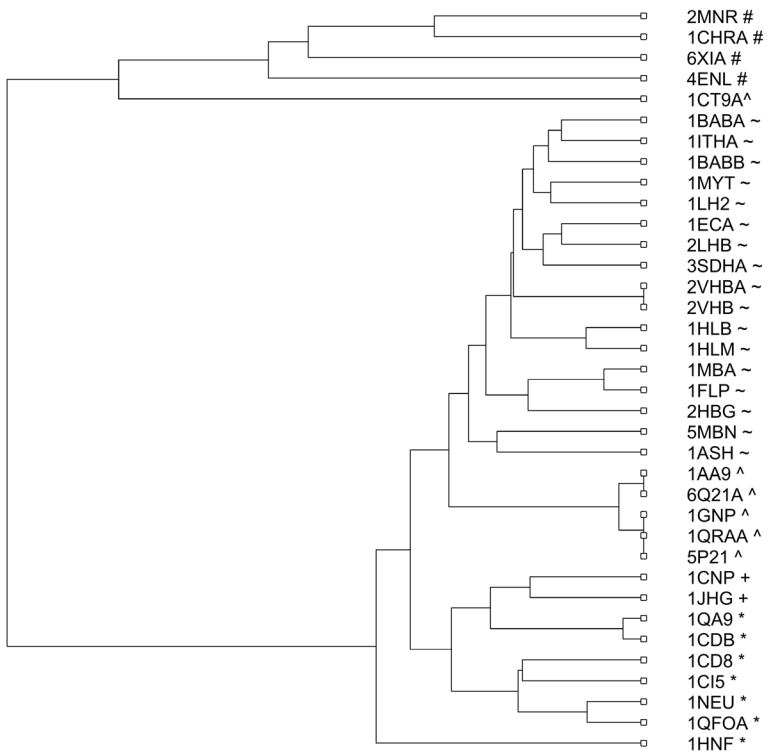

Fig. 2 Phylogenetic tree of the 36 proteins constructed by DCGR

methods erroneously clustered at least three proteins, especially for ClustalW, which erroneously clustered 5 proteins as reported in [43].

Similarity analysis of 50 beta-globin protein sequences

This data set contains 50 beta-globin protein sequences from 50 species studied in [46, 49–53], and the accession numbers have been shown in Additional file 1: Notes 1.2. After extracting features by the method DCGR and reducing the dimensionality using PCA, the *Cosine* distance was used to calculate the distance matrix of 50 beta-globin protein sequences, and the phylogenetic tree was also constructed in Fig. 3.

As shown in Fig. 3, the 50 beta-globin protein sequences are correctly grouped into two clusters corresponding to mammals and non-mammals, respectively. For the mammal cluster, the beta-globin proteins belonging to Carnivora (Black bear, Lesser panda, Giant panda, Coyote, Wolf, Red fox, Dog, Polar bear), Primate (human, grivet, gorilla, langur, gibbon, and chimpanzee), Cetacea (Whale, Dolphin), Bovidae (Sheep, Bison, Buffalo), Proboscidea (Asiatic elephant, African elephant) and Rodentia (Rat, Marmot) are accurately separated and grouped into respective taxonomic classes. In addition, in the branch consisting of Artiodactyla and Perissodactyla, only the Rhinoceros is erroneously clustered. While for the non-mammal cluster, the beta-globin proteins belonging to aves, fish and reptile are also perfectly separated and grouped into respective taxonomic classes. In addition, for the proteins belonging to fishes, the chondrichthyes (Shark) is accurately separated from the actinopterygii (Dragonfish, Cod, Goldfish, Salmon and Catfish) as an independent branch, which is consistent with the known evolutionary relationships.

The phylogenetic trees of other methods [46, 49–53] including ClustalW have also been shown in Additional file 1: Figures S9-S15. After comparison, we found that ClustalW achieves very similar results with our method DCGR, while the other methods performs much worse since even the mammals and non-mammals cannot be correctly separated by the methods in [46, 49–53], and lots of proteins are erroneously clustered by the methods in [46, 51–53].

Similarity analysis of 25 TFs

For this experiment, we select transferrin sequences from 25 vertebrates, which has been well studied by Ford [54]. Their taxonomic information and accession numbers are shown in Additional file 1: Table S3. Similarly processed by DCGR as before, the *Manhattan* distance was used to calculate the distance matrix of the 25 transferrin sequences, and the phylogenetic tree of the 25 TFs was also constructed in Fig. 4.

From Fig. 4, it is easy to find that all the sequences are accurately classified into the fish, amphibian and mammal groups. In the group of mammals, all the sequences

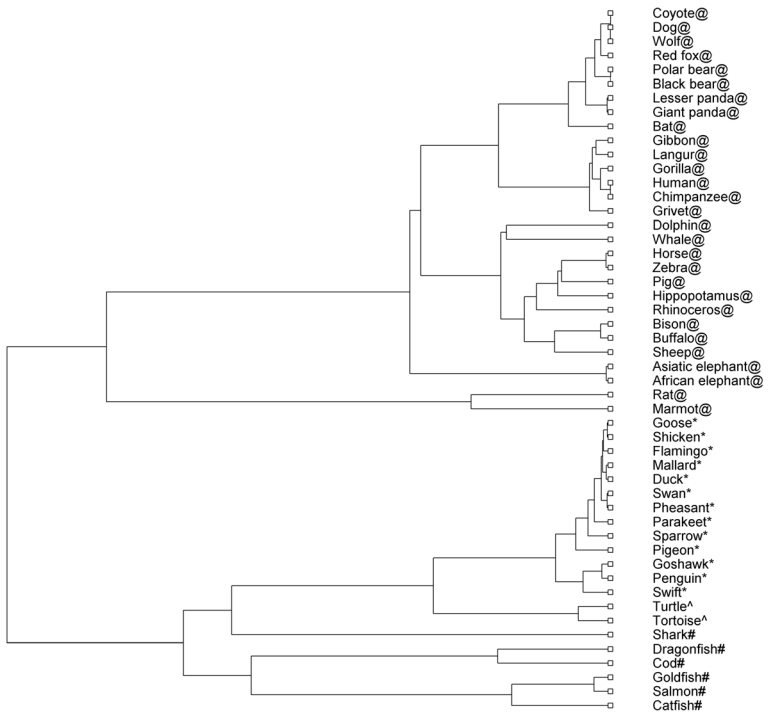

Fig. 3 Phylogenetic tree of the 50 beta-globin protein sequences constructed by DCGR

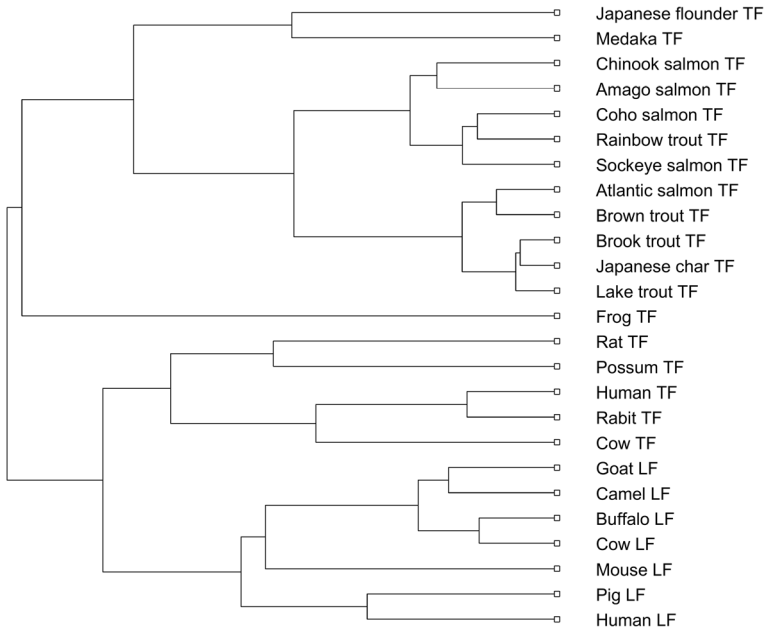

Fig. 4 Phylogenetic tree of the 25 TFs constructed by DCGR

belonging to transferrin (TF) proteins and lactoferrin (LF) proteins are also correctly separated and grouped into respective taxonomic classes. In the group of fishes, all the TFs from Salmonidae are clustered together and form a separate branch. In addition, the TFs belonging to Salmo (Atlantic salmon TF, Brown trout TF), Salvelinus (Lake trout TF, Brook trout TF, Japanese char TF) and Oncorhynchus (Chinook salmon TF, Coho salmon TF, Sockeye salmon TF, Rainbow trout TF, Amago salmon TF) are also correctly clustered and form separate branches, respectively. All these results are completely consistent with known evolutionary relationships. The phylogenetic tree constructed by DCGR is also great consistent with that obtained in [54] (see Additional file 1: Figure S16 for details), which is the most classical result among all the known. However, Possum TF is erroneously clustered in [54], which directly demonstrates that the DCGR is more reliable. For comparison, we also illustrated the phylogenetic tree constructed by ClustalW in Additional file 1: Figure S17, which shows similar results with our method DCGR.

Similarity analysis of 27 AFPs

For the last experiment, the 27 antifreeze protein sequences (AFPs) studied in [43, 52, 55] were used to evaluate the performance of our method. Antifreeze proteins are a class of proteins produced by certain vertebrates, plants, fungi and bacteria that permit their survival in subzero environments by binding to small ice crystals to inhibit growth and recrystallization of ice. The 27 AFPs were selected from Choristoneura fumiferana (CF), *Tenebrio molitor* (TM), *Hypogastrura harveyi* (HH), Dorcus curvidens binodulosus (DCB), Microdera dzhungarica punctipennis (MDP) and *Dendroides Canadensis* (DC), whose taxonomic information and accession numbers are provided in Additional file 1: Table S4. After feature extractions of the 27 AFPs by DCGR, the *standardized Euclidean* distance was used to calculate the distance matrix, and the phylogenetic tree of the 27 AFPs was constructed in Fig. 5. From Fig. 5, it clearly shows that the AFPs of the same species are accurately grouped together. In addition, the HH protein has a far relationship with each of the other 26 AFPs, which is consistent with the result in [56]. However, all the other compared methods [43, 52, 55] including ClustalW cannot accurately group all the proteins into respective taxonomic classes. The phylogenetic trees constructed by these methods have been shown in Additional file 1: Figures S18-S21. For example, ClustalW erroneously divided the TM proteins into two separate groups, while the methods in [43, 52, 55] failed separating the HH protein from the other ones.

We could therefore conclude from all these experiments that our method DCGR demonstrates significant superiority over all the state-of-the-art methods, and it even outperforms the method ClustalW, which is based on sequence alignment.

Importance of the distribution information of points in the CGR image

Applying the distribution information of points in CGR image is a key step in the design of DCGR and makes an essential difference from the other FCGR methods.

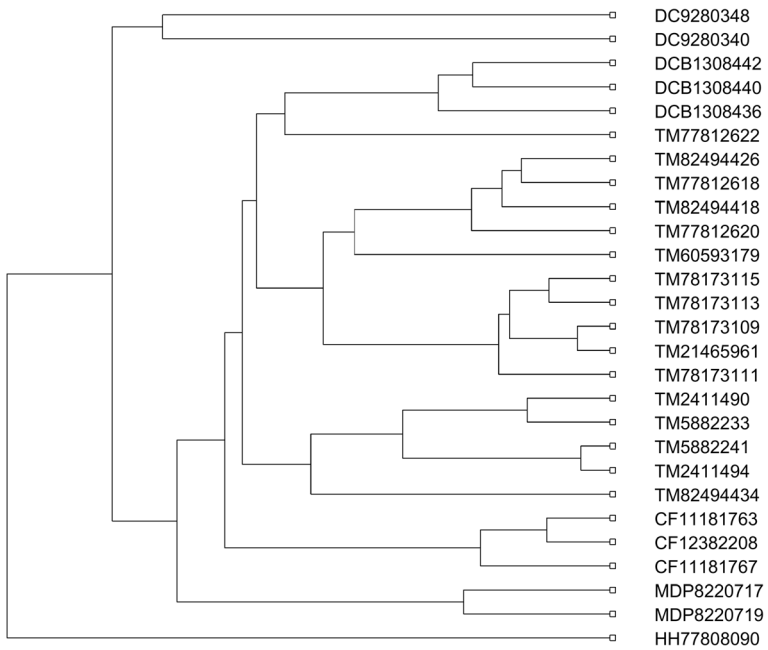

Fig. 5 Phylogenetic tree of the 27 AFPs constructed by DCGR

Traditional FCGR approaches first divide the CGR image into small grids and then take only the point frequency in each grid as numerical characteristics of the sequence without considering the distribution information of the points in each grid as in our method. In order to evaluate the importance of the distribution information of the points in the divided grids, we only took the point frequencies of the four segments as the numerical characteristics of the CGR curve and also used it to construct the phylogenetic trees of the above five data sets, respectively.

After comparison, we found that it performs much worse than DCGR, especially on the second and fifth data sets, whose phylogenetic trees are shown in Figs. 6 and 7, respectively. For the 36 proteins in Fig. 6, the FCGR method without considering the distribution information of points in CGR image separated none of the five protein families from the others, making the phylogenetic tree in quite a mess. For the 27 AFPs in Fig. 7, it erroneously clustered the TM proteins into three branches, and separated the 2 MD proteins in two branches. Similar results could be seen on the other three data sets (see Additional file 1: Figures S22-S24 for details). Therefore, it is easy to conclude that the distribution information of points in the CGR image shows great importance in the method design based on the CGR.

Discussion

Feature extractions of protein sequences play an important role in protein sequence studies, e.g. the predictions of protein functions or protein-protein interactions. Although a great amount of methods have been proposed for extracting features of protein sequences, most of them showed great limits in practical applications. Many studies have showed that the CGR-based strategy would be one of the most useful approaches for protein feature extractions, and the so-called FCGR method is currently the most frequently used method based CGR, however a large amount of useful information, e.g. physicochemical properties of amino acids and the distribution information of points in the CGR image were not taken into consideration in the method design of FCGR.

In this paper, we proposed a new feature extraction method for protein sequences based on the CGR, where two novel techniques are developed in the design of the method DCGR. (1) During the construction of CGR curves, we designed a technique attempting to make full use of the physicochemical properties of amino acids, so the constructed CGR curves contain more useful information, making it more reliable. (2) In the conversion of the CGR curves into numerical characteristics, different from traditional FCGR methods, we opened a new door by integrating the distribution information of points in the CGR image into the method design of DCGR, which is proved quite important and makes the extracted features more efficient.

Compared with previously published methods including ClustalW on five most frequently used data sets, DCGR consistently performs the best. In addition, the method DCGR proposed in this paper could be used not only in the similarity analyses of protein sequences, but also in the areas of investigating protein classification or

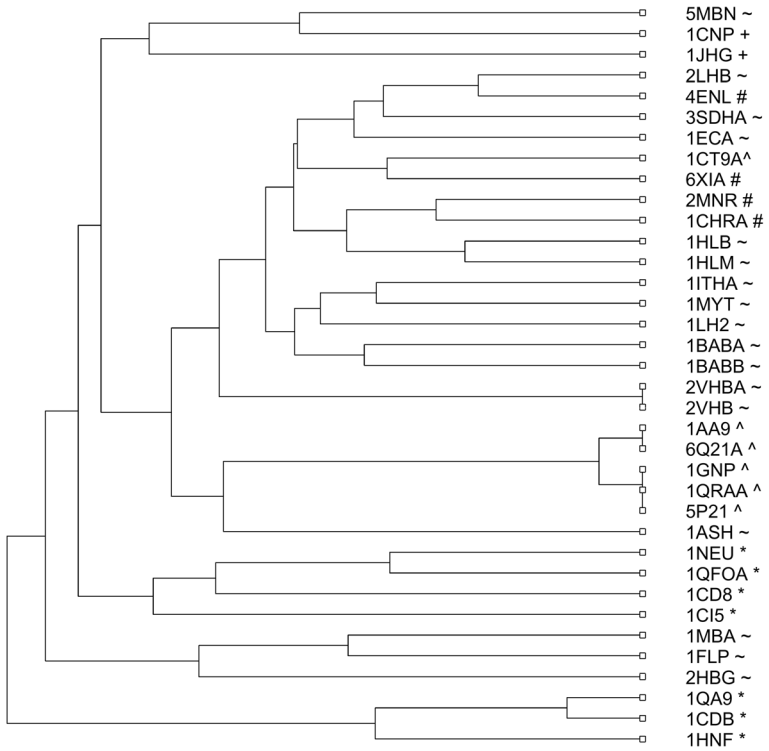

Fig. 6 Phylogenetic tree of the 36 proteins constructed without considering the distribution information of points

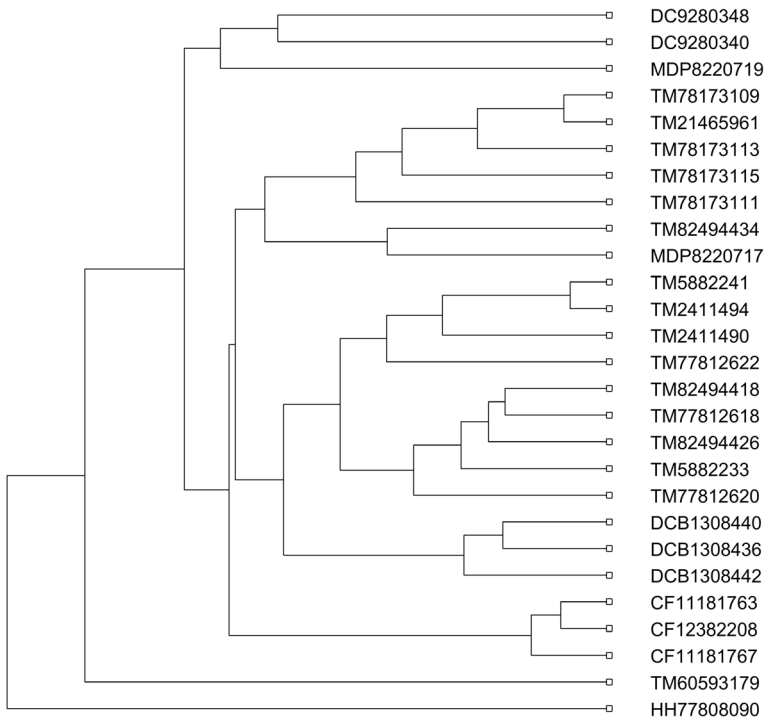

Fig. 7 Phylogenetic tree of the 27 AFPs constructed without considering the distribution information of points

prediction problems in bioinformatics, which will be the topics in our future studies.

Conclusions

We have developed a practically effective method for feature extractions of protein sequences. It is the first CGR-based method by effectively integrating the physicochemical properties of amino acids and the distribution information of points in the CGR image into the method design. Results show that DCGR is currently the most accurate method for protein feature extractions, and demonstrate great potentials for the studies of protein similarity analyses, protein function predictions and protein-protein interactions.

Methods

AAindex database

AAindex is a database of numerical indices representing various physicochemical and biochemical properties of amino acids and amino acid pairs [57, 58]. The latest version is the 9.2 release, which currently contains 566 indices. An amino acid index is a set of 20 numerical values representing any of the different physicochemical properties of the 20 amino acids. Here, we selected 158 indices for the following applications after removing all the redundant ones, in which different amino acids have the same value, and the 158 selected indices have been detailed in Additional file 1: Notes 1.1.

Construction of CGR curves for protein sequences

As did previously, the 4-CGR, 5-CGR or 6-CGR using only 4, 5 or 6 letters to represent protein sequences would result in loss of sequence information, since the amino acids belonging to the same group are considered identical. In order to avoid the loss of sequence information, we developed the DCGR based on 20-CGR mentioned above, where it is a highly challenging task to reasonably locate the 20 amino acids at equal distances on the circumference of a unit circle, as there are up to 20! possible arrangements. In this study, we first developed a novel technique specially to solve the problem of amino acid arrangement by applying the physicochemical properties selected from AAindex database. Then the CGR curves of a protein sequence could be constructed according to the arrangements of the 20 amino acids on the unit circle.

Arranging the 20 amino acids on the circumference of a unit circle

In order to fully use the physicochemical properties of the amino acids, we first sort the 20 amino acids according to their physicochemical indices in ascending order. Then the 20 amino acids are arranged in order on the circumference of a unit circle by the following equation,

$$\phi(X_i) = \left(\cos\frac{2\pi i}{20},\ \sin\frac{2\pi i}{20} \right), i = 1, 2, \cdots, 20 \qquad (1)$$

where X_i represents each of the 20 amino acids.

Building CGR curves for protein sequences

Given a protein sequence S with N amino acids $S = s_1 s_{2\ldots}s_{\mathrm{N}}$, the CGR curve is constructed by successively connecting N points corresponding to the N amino acids, the coordinate of which are determined as follows. The first point is specified as the midpoint of the center of the unit circle and the point of the circumference corresponding to the first amino acid s_1. For the i-th amino acid s_i, its point coordinate is defined as the midpoint of the (i-1)-th point and the point of the circumference corresponding to the amino acid s_{i}. In detail, the iterative procedure can be formulated as:

$$\psi(s_i) = \frac{1}{2}(\psi(s_{i-1}) + \phi(s_i)), i = 1, 2, \cdots, N \qquad (2)$$

where $\psi(s_i)$ represents the coordinate of the point corresponding to the i-th amino acid s_i, and $\psi(s_0)$ is set to be (0, 0).

Corresponding to each of the 158 selected physicochemical properties, we can obtain an exclusive arrangement of 20 amino acids on the circumference of a unit circle, and then a CGR curve for a protein sequence. Thus, 158 intrinsically different CGR curves could be constructed for each protein sequence corresponding to the 158 physicochemical properties of amino acids.

Conversion of CGR curves into numerical characteristics

After obtaining 158 CGR curves for each protein sequence, another challenging task is to effectively convert the CGR curves into numerical characteristics, which could then be used for similarity analysis among protein sequences. In this study, we developed a new method for extracting numerical characteristics from CGR curves as follows.

Given a protein sequence S, we can obtain 158 different CGR curves falling in a unit circle. In order to extract features from protein sequence, for each of the 158 CGR curves, we first split the unit circle into four segments according to the four quadrants. Then, we compute pairwise distances between points in each segment and obtain four distance matrices for a CGR curve. By computing their leading eigenvalues, we obtain a 4-dimensional vector which is taken as the numerical characteristics of the CGR curve. All of the numerical characteristics of 158 CGR curves are integrated into a 632-dimensional vector which is taken as the feature vector of the protein sequence.

Given a data set consisting of N protein sequences, we can obtain an $N \times 632$ feature matrix, each row of which corresponds to a feature vector of a protein sequence.

Since the dimension of the feature vectors is very high, there may be redundancies and noises in them. We use the Principal Component Analysis (PCA) to reduce the dimensionality of the feature vectors. The reduced feature vectors are then applied to analyze the similarity of protein sequences.

Additional file

Additional file 1: This file contains supplementary notes, figures and tables. (PDF 1238 kb)

Abbreviation
CGR: Chaos Game Representation

Acknowledgements
Not applicable.

Authors' contributions
JL and GL conceived and designed the approach. ZM and TY implemented the software. ZM and TY performed data analysis. ZM, EQ and JL wrote the manuscript. GL supervised and revised the manuscript. All authors approved the final version of this manuscript.

Funding
This work was supported by the National Natural Science Foundation of China with code 61801265, 61432010, 61771009, the Shandong Provincial Natural Science Foundation, China with code ZR2018PA001, and Research Fundamental Capacity Improvement Project for Middle Age and Youth Teachers of Guangxi Universities with code 2019KY0078. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials
See Additional file 1 for the availability of the tested protein sequences, and DCGR is a free, open-source package available from https://sourceforge.net/projects/transcriptomeassembly/files/Feature%20Extraction.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]School of Mathematics, Shandong University, Jinan 250100, Shandong Province, China. [2]College of Mathematics and Statistics, Guangxi Normal University, Guilin 541001, China.

Received: 1 February 2019 Accepted: 10 June 2019
Published online: 20 June 2019

References
1. Rigden DJ. From protein structure to function in bioinformatics. New York: Springer-verlag; 2009.
2. Qi Z, Li K, Ma J, Yao Y, Liu L. Novel method of 3-dimensional graphical representation for proteins and its application. Evol Bioinforma. 2018;14:1–8.
3. Li C, Zhao J, Wang C, Yao Y. Protein sequence comparison and DNA-binding protein identification with generalized PseAAC and graphical representation. Comb Chem High Throughput Screen. 2018;21:100–10.
4. Mehri M, Fatemeh A, Vahid Z. A novel graphical representation and similarity analysis of protein sequences based on physicochemical properties. Physica A. 2018;510:477–85.
5. Mu Z, Li G, Wu H, Qi X. 3D-PAF curve: a novel graphical representation of protein sequences for similarity analysis. Match Commun Math Comput Chem. 2016;75:447–62.
6. Huang G, Hu J. Similarity/dissimilarity analysis of protein sequences by a new graphical representation. Curr Bioinforma. 2013;8:539–44.
7. Li Z, Geng C, He P, Yao Y. A novel method of 3D graphical representation and similarity analysis for proteins. Match Commun Math Comput Chem. 2014;71:213–26.
8. el Maaty MIA, Abo-Elkhier MM, Elwahaab MAA. 3D graphical representation of protein sequences and their statistical characterization. Physica A. 2010; 389:4668–76.
9. Gupta MK, Niyogi R, Misra M. A 2D graphical representation of protein sequence and their similarity analysis with probabilistic method. Match Commun Math Comput Chem. 2014;72:519–32.
10. He P, Li X, Yang J, Wang J. A novel descriptor for protein similarity analysis. Match Commun Math Comput Chem. 2011;65:445–58.
11. Yu JF, Sun X, WANG JH. A novel 2D graphical representation of protein sequence based on individual amino acid. Int J Quantum Chem. 2011; 111:2835–43.
12. Liu Y, Li D, Lu K, Jiao Y, He P, Curve P-H. A graphical representation of protein sequences for similarities analysis, MATCH Commun. Math Comput Chem. 2013;70:451–66.
13. Wu ZC, Xiao X, Chou KC. 2D-MH: a web-server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids. J Theor Biol. 2010;267:29–34.
14. Ma T, Liu Y, Dai Q, Yao Y, He P. A graphical representation of protein based on a novel iterated function system. Physica A. 2014;403:21–8.
15. Wen J, Zhang YY. A 2D graphical representation of protein sequence and its numerical characterization. Chem Phys Lett. 2009;476:281–6.
16. Bai F, Wang T. On graphical and numerical representation of protein sequences. J Biomol Struct Dyn. 2006;23:537–45.
17. el Maaty MIA, Abo-Elkhier MM, Elwahaab MAA. Representation of protein sequences on latitude-like circles and longitude-like semi-circles. Chem Phys Lett. 2010;493:386–91.
18. Li C, Xing L, Wang X. 2-D graphical representation of protein sequences and its application to coronavirus phylogeny. BMB Rep. 2008;41:217–22.
19. Yao Y, Yan S, Han J, Dai Q, He P. A novel descriptor of protein sequences and its application. J Theor Biol. 2014;347:109–17.
20. Liao B, Liao B, Lu X, Cao Z. A novel graphical representation of protein sequences and its application. J Comput Chem. 2011;32:2539–44.
21. Li D, Wang J, Li C. New 3-D graphical representation of protein sequences and its application. Chin J Bioinf. 2009;7:60–3.
22. Jeffrey H. Chaos game representation of gene structure. Nucleic Acids Res. 1990;18:2163–70.
23. Joseph J, Sasikumar R. Chaos game representation for comparision of whole genomes. BMC Bioinf. 2006;7:243–52.
24. Randić M, Zupan J. Highly compact 2D graphical representation of DNA sequences. SAR QSAR Environ Res. 2004;15:191–205.
25. Nair N, Nair A. Combined classifier for unknown genome classification using chaos game representation features. https://doi.org/10.1145/1722024.1722065.
26. Adetiba E, Badejo J, Thakur S, Matthews V, Adebiyi M, Adebiyi E. Experimental investigation of frequency chaos game representation for in silico and accurate classification of viral pathogens from genomic sequences. https://doi.org/10.1007/978-3-319-56148-6_13.
27. Tanchotsrinon W, Lursinsap C, Poovorawan Y. An Efficient Prediction of HPV Genotypes from Partial Coding Sequences by Chaos Game Representation and Fuzzy k-Nearest Neighbor Technique. https://doi.org/10.2174/1574893611666161611101120.
28. Tanchotsrinon W, Lursinsap C, Poovorawan Y. A high performance prediction of HPV genotypes by chaos game representation and singular value decomposition. https://doi.org/10.1186/s12859-015-0493-4.
29. Fiser A, Tusnády G, Simon I. Chaos game representation of protein structures. J Mol Graph. 1994;12:302–4.
30. Randić M, Butina D, Zupan J. Novel 2-D graphical representation of proteins. Chem Phys Lett. 2006;419:528–32.
31. He P, Zhang Y, Yao Y, Tang Y, Nan X. The graphical representation of protein sequences based on the physicochemical properties and its applications. J Comput Chem. 2010;31:2136–42.
32. Yu Z, Anh V, Lau K. Chaos game representation of protein sequences based on the detailed HP model and their multifractal and correlation analyses. J Theor Biol. 2004;226:341–8.

33. Manikandakumar K, Gokulraj K, Muthukumaran S, Srikumar R. Graphical representation of protein sequences by CGR: analysis of pentagon and hexagon structures. https://doi.org/10.5829/idosi.mejsr.2013.13.6.2344.
34. Hu X, Xia J, Niu X, Ma X. Chaos game representation for discriminating thermophilic from mesophilic protein sequences. https://doi.org/10.1109/ICBBE.2009.5162487.
35. Li N, Shi F, Niu X, Xia J. A novel method to reconstruct phylogeny tree based on the chaos game representation. J Biomed Sci Eng. 2009;2:582–6.
36. Niu X, Shi F, Hu X, Xia J, Li N. Predicting the protein solubility by integrating chaos games representation and entropy in information theory. Expert Syst Appl. 2014;41:1672–9.
37. Niu X, Hu X, Shi F, Xia J. Predicting protein solubility by the general form of Chou's pseudo amino acid composition: approached from chaos game representation and fractal dimension. Protein Pept Lett. 2012;19:940–8.
38. Wang H, Wu P. Prediction of RNA-protein interactions using conjoint triad feature and chaos game representation. Bioengineered. 2018;9:242–51.
39. Randić M. 2-D graphical representation of proteins based on physico-chemical properties of amino acids. Chem Phys Lett. 2007;440:291–5.
40. He P. A new graphical representation of similarity/dissimilarity studies of protein sequences. SAR QSAR Environ Res. 2010;21:571–80.
41. Basu S, Pan A, Dutta C, Das J. Chaos game representation of proteins. J Mol Graphics Modell. 1997;15:279–89.
42. Wang Y, Hill K, Singh S, Kari L. The spectrum of genomic signatures: from dinucleotides to chaos game representation. Gene. 2005;346:173–8.
43. Wu H, Zhang Y, Chen W, Mu Z. Comparative analysis of protein primary sequences with graph energy. Physica A. 2015;437:249–62.
44. Zhang S, Yang L, Wang T. Use of information discrepancy measure to compare protein secondary structures. J Mol Struct Theochem. 2009;909:102–6.
45. Krasnogor N, Pelta DA. Measuring the similarity of protein structures by means of the universal similarity metric. Bioinformatics. 2004;20:1015–21.
46. Xu C, Sun D, Liu S, Zhang Y. Protein sequence analysis by incorporating modified chaos game and physicochemical properties into Chou's general pseudo amino acid composition. J Theor Biol. 2016;406:105–15.
47. Mu Z, Wu J, Zhang Y. A novel method for similarity/dissimilarity analysis of protein sequences. Physica A. 2013;392(24):6361–6.
48. Wang Y, Wu LY, Zhang JH, Zhan ZW, Zhang XS, Chen L. Evaluating protein similarity from coarse structures. IEEE/ACM Trans Comput Biol Bioinf. 2009;6:583–93.
49. Yu C, He R, Yau S. Protein sequence comparison based on K-string dictionary. Gene. 2013;529:250–6.
50. Tian K, Yang X, Kong Q, Yin C, He R, Yau S. Two dimensional Yau-Hausdorff distance with applications on comparison of DNA and protein sequences. https://doi.org/10.1371/journal.pone.0136577.
51. Yau S, Yu C, He R. A protein map and its application. Dna Cell Biol. 2008;27:241–50.
52. Yu L, Zhang Y, Gutman I, Shi Y, Dehmer M. Protein sequence comparison based on physicochemical properties and the position-feature energy matrix. https://doi.org/10.1038/srep46787.
53. Wan X, Zhao X, Yau S. An information-based network approach for protein classification. https://doi.org/10.1371/journal.pone.0174386.
54. Ford M. Molecular evolution of transferrin: evidence for positive selection in salmonids. Mol Biol Evol. 2001;18:639–47.
55. Zhang Y. A new model of amino acids evolution, evolution index of amino acids and its application in graphical representation of protein sequences. Chem Phys Lett. 2010;497:223–8.
56. Lin F, Laurie A, Robert L, Peter L. Structural modeling of snow flea antifreeze protein. Biophys J. 2007;92:1717–23.
57. Nakai K, Kidera A, Kanehisa M. Cluster analysis of amino acid indices for prediction of protein structure and function. Protein Eng. 1988;2:93–100.
58. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. AAindex: amino acid index database, progress report 2008. Nucleic Acids Res. 2008;36:D202–5.

Publisher's Note