Journal of Theoretical Biology 331 (2013) 48–53

Contents lists available at SciVerse ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Using protein granularity to extract the protein sequence features

Zhi-Xin Liu [a,c], Song-lei Liu [b], Hong-Qiang Yang [c,*], Li-Hua Bao [d]

[a] Department of Applied Physics, Shandong Agricultural University, Taian, Shandong 271018, China
[b] College of Life Sciences, Tsinghua University, Beijing 100084,China
[c] State Key Laboratory of Crop Biology, National Apple Engineering Technology Center, College of Horticultural Science and Engineering, Shandong Agricultural University, Taian, Shandong 271018, China
[d] College of Chemistry and Material Science , Shandong Agricultural University, Taian, Shandong 271018, China

ARTICLE INFO

Article history:
Received 9 December 2012
Received in revised form
16 April 2013
Accepted 18 April 2013
Available online 26 April 2013

Keywords:
Amino acid composition
Length effect
Structure class prediction
Support vector machine
Increment

ABSTRACT

The feature extraction of protein sequences is a challenging problem. It might need a lot of theoretical and practical knowledge from many fields. The difficulty would increase when investigators extract the features solely from protein sequences. In this paper, we present a method of protein granularity. The concepts of protein granularity, granularity order, granularity bound, granularity limit, and granularity increment are given respectively. The protein granularity can dig out the useful information solely from protein sequences. We provide an approach to construct the feature vectors. The feature vectors include the amino acid composition information, the sequence-order information, the same amino acid 'neighbor' information, and the sequence length information. Hence, the feature vectors can better represent protein sequences. Our feature extraction method does obviously consider the protein sequence length effects. An experiment of the protein structure class prediction was carried out. The prediction achieved 96.6% overall accuracy, and the success rate for each subset is all-α 92.3%, all-β 100%, α/β 100%, $\alpha+\beta$ 93.5%, respectively. The last three success rates for subsets are equal to the best success rates in the published literatures. The overall accuracy of PG-SVM prediction is the second best result only having one protein prediction error difference with the first best result. The theoretical and experimental results demonstrate the application of protein granularity succeeds in the feature extraction of protein sequences.

1. Introduction

The amino acid sequence plays a crucial role in the determination of three-dimensional structure of a protein molecule (Anfinsen, 1973; Baker, 2000). The number of newly found protein sequences has increased greatly in the post genomic era. Most of them haven't any other or enough useful information on the three-dimensional structure and function. If investigators want to know these kinds of information, they can conduct a series of biological experiments, use the bioinformatics methods to predict it, or carry out biophysical analysis methods. The biological experiments are good ways to finish these tasks, but the bioinformatics methods and biophysical analysis methods are proper in some situations. For instance, Cherstvy et al. propose the chargeable group concepts to analyze reaching and recognizing the targets about the protein-DNA interactions (Cherstvy et al., 2008; Cherstvy, 2009).

Many feature extraction methods have been proposed and are widely used in the protein predictions, such as utilizing amino acid composition(AAC) to predict protein cellular distribution (Du and Li, 2006; Zhou and Doctor, 2003), utilizing amino acid composition to predict protein structural class (Feng et al., 2005; Jahandideh et al., 2007), utilizing dipeptide composition to predict protein subcellular locations (Huang and Li, 2004), utilizing polypeptide to predict protein structural class (Luo et al, 2002; Sun and Huang, 2006), utilizing the increment of diversity to predict the subcellular location of apoptosis proteins (Chen and Li, 2007), utilizing the amino acid hydrophobicity to predict protein structural class (Qui et al., 2008), utilizing grouped weight to predict apoptosis protein subcellular localization (Zhang et al., 2006), and utilizing physicochemical composition features to predict subnuclear localization (Huang et al., 2007). The pseudo amino acid composition (PseAAC) is one of the frequently and widely used feature extraction methods of protein sequences (Chou, 2009, 2011).

In this paper, we give several concepts about protein granularity. Then we use protein granularity to extract the protein sequence features and provide an approach to construct the feature vectors of protein sequences. An experiment of the protein structure class prediction was carried out and then we analyze the results.

2. Protein sequence and its granularity

Granularity idea originates from the concepts of coarse-grained description and grouping in physics. Using the idea to investigate

* Corresponding author. Tel.: +86 0538 8249304.
E-mail address: zxliu@sdau.edu.cn (H.-Q. Yang).

0022-5193/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jtbi.2013.04.019

the protein sequence, we have a new method of protein granularity for the feature extraction of protein sequences.

2.1. Protein granularity

We first define a concept: granularity of the amino acid sequence of protein, or call it protein granularity (PG).

We have a set $B=\{A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y\}$, where A represents alanine, C represents cysteine, … , Y represents tyrosine. Each element in set B belongs to one of the 20 native amino acids. Suppose we also have a set $Z=\{z_1,z_2,...,z_m\}$, and Z is a subset of B , where $1\le m\le 20$, Now, let a set $X=\{x_1,x_2,...,x_n\}$, where n is a positive integer. If the elements in X and Z are ordered, and suppose they satisfy

$$x_1 < x_2 < \cdots < x_n, z_1 < z_2 < \cdots < z_m.$$

Then map $f : X\to Z$ corresponds to a n-letter array $f(x_1)f(x_2)\cdots f(x_n)(f(x_i)\in Z)$. We also call the array a word composed with n letters which is randomly selected from the set Z. When $f(x_1)\le f(x_2)\le\cdots\le f(x_n)$, we call $f(x_1)f(x_2)\cdots f(x_n)$a protein granularity.

2.2. Protein granularity order

In the above set $X=\{x_1,x_2,...,x_n\}$, we call n the protein granularity order. If there is only one element in the set X, we define granularity order is the first-order. If there are two elements in the set X, we define granularity order is the second-order, and so on.

With a protein sequence, we not only can get the number (c^n_{type}) of total granularity types at the n^{th}-order level, but also can get the value (c^n_{same}) of frequency of a granularity at the same order level. Let's see a protein domain sequence (PDB: 1RDH_A) fragment: a part of HIV RNase H, 11 amino acids, "PFHGYQLEKEP" .

When we take the 1st-order granularity from the fragment, we only get one letter along the sequence. If the letter appears for the first time, we get a new 1st-order granularity, and the frequency of the new granularity is 1. If the letter appears for the second time, we add 1 to the value of the corresponding granularity's frequency, and so on.

When we take the 2nd-order granularity from the fragment, we get two letters along the sequence. After reordering the two letters according to the alphabet, we get a 2nd-order granularity. If the granularity appears for the first time, the frequency of the granularity is 1. If the granularity appears for the second time, we add 1 to the value of the corresponding granularity's frequency, and so on. To get 3rd-order granularities, 4th-order granularities, etc. the steps are similar with the above steps, and all of the granularities are finally obtained (Table 1).

From the process of getting the granularities, we can deduce the sequence-order effects are already better considered and the information of sequence-order is included into the granularities. Obviously, the granularities also include AAC information.

From table 1, we find a new phenomenon. Amount of the 2nd-order "EK" is 2, being from "EKE" in the fragment. It is because the second "E" is a 'neighbor' of the previous "E", and they are separated by one residue. Amount of 10th-order "EEFGHKLPQY" is 2, being from "PFHGYQLEKEP". It is because the second "P" is a 'neighbor' to the previous "P", and they are separated by nine residues. Hence, we deduce this kind of granularities can reflect the same amino acid 'neighbor' effects in the sequence.

What is the general performance of the protein sequences about protein granularity? Let's see an example: CAS1A_XENLA (Swiss-Prot: P55865,): CAS1A_XENLA has 386 amino acids.

We take the 2nd-order granularity and 3rd-order granularity from the sequence respectively. The meaning of the frequency in the Fig. 1, Fig. 2 is the number of one granularity which appears in the fragment. Fig. 1 shows that the frequency of 3rd-order granularities ranges from 1 to 4. The granularity frequency distribution also has some differences 'along' the protein sequence, but also not one frequency is overwhelmingly higher than others. In our computation, we find the frequency of 2nd-order granularities ranges from 1 to 8 and the number of total granularity type is 153. Compared with the frequency of 2nd-order granularities, the maximum of frequency of 3rd-order granularities becomes smaller. The number of total granularity type changes from 153 to 309. We further study many other protein sequences, and find results are similar to the above results. And we also find the long protein sequences usually have the large numbers of total granularity types at the same order level.

We divided CAS1A_XENLA sequence into two equal length fragments, and take the 2nd-order granularity from each of fragments. After the granularities are reordered with alphabet, we have the granularity frequency result (Fig. 2). The Fig. 2 shows the frequency maximums of the two fragments equal 5. The discretion degree of granularity frequency of two fragments is similar, and not one frequency is overwhelmingly larger than others. We also find some granularities just appear in one fragment. The number of total granularity types of the first fragment is 105, and the number of total granularity types of the second fragment is 120. Both of them are smaller than the number of total granularity types of the whole sequence (153).

2.3. Protein granularity bound

Theorem. Give a n-element ordered set $X=\{x_1, x_2, ..., x_n\}$ $(x_1 < x_2 < \cdots < x_n)$, where n is a positive integer. And have 20-element ordered set $Z=\{A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V, W,Y\}$. Each element in set Z belongs to one of the 20 native amino acids. The number of total granularity types at n^{th}-order level, the granularity corresponding to map $f : X\to Z$, equals to the combinatorial number of repeatable selection of n amino acids from 20 amino acids. And the number is

$$|f(x_1)f(x_2)\cdots f(x_n)| = C^n_{20+n-1}, \quad (1)$$

where $|\{\bullet\}|$ is the number of the set's elements, C^n_{20+n-1} is combinatorial number.

We call the number $c^n_{upper} = |\{f(x_1)f(x_2)\cdots f(x_n)\}|$ the n^{th}-order protein granularity upper bound. The 1st-order upper bound c^1_{upper} is 20; the 2nd-order upper bound c^2_{upper} is 210, the 3rd-order upper bound c^3_{upper} is 1540 and so on. Obviously, the n^{th}-order granularity lower bound $c^n_{lower} = 0$.

2.4. Granularity increment

We further have, to a protein sequence, the ratio (r^n_{in})which is the number (c^n_{type}) of total granularity types at the n^{th}-order level to n^{th}-order granularity upper bound c^n_{upper}. We call r^n_{in} the granularity increment at the n^{th}-order level.

$$r^n_{in} = c^n_{type}/c^n_{upper}. \quad (2)$$

2.5. Granularity limit of a protein sequence

Given a protein sequence, we can get a series of numbers of total granularity types at different order level. Further studies show there is a maximum in the numbers, and we call the maximum the granularity limit of the protein sequence. Let's see an example: POLG_HCVEV (Swiss-Prot: O39928). POLG_HCVEV has 3014 amino acids. We select POLG_HCVEV as one example because the long protein sequence can make the following results be shown very clearly (Fig. 3). Our computation show that the granularity limit of POLG_HCVEV protein sequence is 2807 while the granularity order is 15 (see the point "∗" in Fig. 3). From Fig. 3,