Methods 69 (2014) 207–212

Contents lists available at ScienceDirect

Methods

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

SFAPS: An R package for structure/function analysis of protein sequences based on informational spectrum method

Su-Ping Deng [a], De-Shuang Huang [a,b,*]

[a] Machine Learning and Systems Biology Lab, College of Electronics and Information Engineering, Tongji University, 4800 Caoan Road, Dianxin Building, Shanghai 201804, China
[b] College of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou 450002, China

ARTICLE INFO

Article history:
Received 20 March 2014
Revised 2 August 2014
Accepted 6 August 2014
Available online 15 August 2014

Keywords:
R package
Structure/function analysis
Protein sequence
Informational spectrum method

ABSTRACT

The R package SFAPS has been developed for structure/function analysis of protein sequences based on information spectrum method. The informational spectrum method employs the electron–ion interaction potential parameter as the numerical representation for the protein sequence, and obtains the characteristic frequency of a particular protein interaction after computing the Discrete Fourier Transform for protein sequences. The informational spectrum method is often used to analyze protein sequences, so we developed this software tool, which is implemented as an add-on package to the freely available and widely used statistical language R. Our package is distributed as open source code for Linux, Unix and Microsoft Windows. It is released under the GNU General Public License. The R package along with its source code and additional material are freely available at http://mlsbl.tongji.edu.cn/DBdownload.asp.

1. Introduction

Proteins are one of the most complex and varied classes of macromolecules found in a cell. They have many kinds of functions. As enzymes, they catalyze countless chemical reactions. They are active as carriers and storage molecules in muscle contraction and in mechanical support. As antibodies, they are responsible for immune protection. As receptors in the nervous system, they are responsible for the generation and transmission of nerve impulses. Proteins are polymers built up from amino acids. The great diversity and versatility of protein sequences are derived from the properties of 20 different amino acid side chains. Proteins can express their biological functions only by achieving a certain active conformation that is the so-called three-dimensional (3D) structure. Apparently, the particular function of a given protein and its active 3D structure are basically determined by its sequence of amino acids. The proteins' biological function is hidden in the proteins' primary structure, i.e., the sequence of amino acids. There have been many attempts to discover the main principles controlling the functional behaviors of proteins. Typical approaches are either homology characterizing of specific features of the primary and secondary structure of proteins or molecular modeling of the proteins' 3D structure. Although such approaches provide a significant insight into the proteins' structure and active sites, they do not provide enough knowledge about informational, structural and physicochemical parameters, which are important for the selectivity of protein interactions that can be useful for the *de novo* design of peptides or proteins analogous to the desired biological activity [1,2].

The amino acid sequence of a protein is a valuable source to investigate its function, structure and history [3]. The sequence of a protein can be compared with other known sequences to decide whether significant similarities exist. The search for similarity between a new sequenced protein and millions of previously sequenced ones takes a few seconds using computers. If the newly isolated protein is a member of an established class of proteins, we can infer information about the protein's structure and function. The phylogenic relationships among species can be inferred based on the differences in their protein sequences. Assuming that the mutation rate of proteins is constant, the analysis of sequences of homologous proteins from different species is able to give information when every two species diverged. Sequences provide a basis for preparing antibodies specific for a protein of interest. Parts of an amino acid sequence can be extracted as an antibody when injected into a mouse or rabbit. Amino acid sequences are valuable for making DNA probes used to encode its proteins. By knowing the primary structure, it permits the use of reverse genetics. The DNA sequences, which correspond to part of an amino acid sequence, can be constructed on the basis of genetic code. These DNA

* Corresponding author at: Machine Learning and Systems Biology Lab, College of Electronics and Information Engineering, Tongji University, 4800 Caoan Road, Dianxin Building, Shanghai 201804, China.
E-mail address: dshuang@tongji.edu.cn (D.-S. Huang).

http://dx.doi.org/10.1016/j.ymeth.2014.08.004
1046-2023/

sequences can be used as probes to isolate the gene encoding the protein so that the entire sequence can be determined. The gene in turn can provide information about the physiological regulation of the protein.

According to the electron–ion interaction potential concept [4], the electrons along a protein molecule are considered to be delocalized. The charges moving along the protein backbone induce transient polarization of the side groups, resulting in attraction and repulsion between some parts of the molecule and oscillation of the molecule as a whole. These oscillations can be propagated through polar media (water) at large distances (100–1000A) and interfere with oscillations of other molecules [5–7]. As demonstrated by Fröhlich [6], between two molecules encompassed in their oscillations, the same frequency component can establish the resonant interaction which results in highly specific attractive forces between these molecules. It is also proposed that biological macromolecules can attract small molecules at large distance and induce their "passive" oscillations. This long-distance interaction directly influences a number of productive collisions between interacting biochemical molecules and their kinetics. The characteristics of these oscillations are determined by the electronic properties of the "building blocks" (amino acids and nucleotides) and their distribution along the sequence. It has been demonstrated that the electron–ion interaction potential (EIIP) [4,8,9] of amino acids and nucleotides represents an essential physical property determining characteristics of these molecular oscillations.

The informational spectrum method (ISM) is one of methods to process protein (or nucleotide) sequences. In terms of ISM, the protein (or nucleotide) sequences are transformed into signals by assignment of numerical values of each element (amino acid or nucleotide). These values correspond to EIIP determining electronic properties of amino acids and nucleotides. The obtained signal is then decomposed in periodical function by Fourier transformation. The result is series of frequencies and their amplitudes. The obtained frequencies correspond to the distribution of structural motifs with defined physicochemical characteristics that are responsible for biological function of the sequence. When comparing proteins which share the same biological or biochemical function, the technique allows the detection of code/frequency pairs which are specific for their common biological properties [10]. The ISM is insensitive to the location of the motifs, and thus, does not require previous alignment of the sequences. The ISM has been successfully applied in prediction of biological function of novel proteins, structure/function analysis of different protein and DNA sequences and *de novo* design of biologically active peptides [10].

2. Methods and implementation

2.1. Informational spectrum method

Before applying discrete Fourier transform to a protein sequence, the signal sequence of the protein needs to be mapped to a numerical sequence first. Here, the mapping approach is based on the electron–ion interaction potential (EIIP) [11]. The values of EIIP represent the main energy term of valence electrons, which are essential physical parameters determining the long-range properties of biological molecules. The EIIP can be determined for organic molecules by simple equations derived from the "general model pseudo-potential" [4,8] and are presented in Table 1.

The physical and mathematical basis of ISM was described in details elsewhere [2,12] and here we will only present this bioinformatics method in brief. A sequence of N residues is represented as a linear array of N terms, with each term given a weight. In this way the alphabetic code is transformed into a sequence of numbers. The obtained numerical sequence, representing the primary structure of protein, is then subjected to a Fast Fourier transformation (FFT), which is defined as follows:

Table 1
The EIIP values of amino acids.

Amino acid	EIIP
Leu	0
Ile	0
Asn	0.0036
Gly	0.0050
Val	0.0057
Glu	0.0058
Pro	0.0198
His	0.0242
Lys	0.0371
Ala	0.0373
Tyr	0.0516
Trp	0.0548
Gln	0.0761
Met	0.0823
Ser	0.0829
Cys	0.0829
Thr	0.0941
Phe	0.0946
Arg	0.0959
Asp	0.1263

$$X(n) = \sum_{m=0}^{N-1} x(m) e^{-j(2\pi nm)/N}, \quad n = 1, 2, \ldots, N/2 \tag{1}$$

where $x(m)$ is the mth member of a given numerical series, N is the total number of points in this series, and $X(n)$ are FFT coefficients. These coefficients describe the amplitude, phase and frequency of sinusoids, which comprised the original signal. The complete information about the original sequence is contained in both spectral functions. A FFT is an algorithm to compute the discrete Fourier transform (DFT) and its inverse. Fourier analysis converts time (or space) to frequency and vice versa; an FFT rapidly computes such transformations by factorizing the DFT matrix into a product of sparse (mostly zero) factors [13]. As a result, fast Fourier transforms are widely used for many applications in engineering, science, and mathematics. Fast Fourier transforms have been described as "the most important numerical algorithm[s] of our lifetime" [14]. However, for protein analysis, relevant information is presented in an energy density spectrum [12,15], which is defined as follows:

$$S(n) = X(n)X^*(n) = X^2(n), \quad n = 1, 2, \ldots, N/2 \tag{2}$$

where $X(n)$ are the FFT coefficients of the series and $X^*(n)$ are complex conjugate FFT coefficients of the series.

Numerical series obtained in this way are then analyzed by digital signal analysis methods in order to extract information relevant to the biological function. The original numerical sequences are transformed to the frequency domain using the Fast Fourier Transform (FFT). In this way, sequences are analyzed as discrete signals. It is assumed that their points are equidistant. For further numerical analysis, the distance between points in these numerical sequences is set at an arbitrary value. And for the convenience of computation, we set $d = 1$ s. Then the maximal frequency in a spectrum defined in this way is $F = 1/2d = 0.5$ Hz. The frequency range is independent of the total number of points in the sequence. The total number of points in a sequence only influences the resolution of spectrum. The resolution of the N-point sequence is $1/N$. The n-th point in the spectral function corresponds to a frequency $f(n) = nf = n/N$. Thus, the initial information defined by the sequence of amino acids can now be presented in the form of an informational spectrum (IS), representing a series of frequencies and their amplitudes.