BioSystems 168 (2018) 45–54

Contents lists available at ScienceDirect

BioSystems

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

Amino acid content of beta strands and alpha helices depends on their flanking secondary structure elements

Vladislav Victorovich Khrustalev[a,*], Tatyana Aleksandrovna Khrustaleva[b], Victor Vitoldovich Poboinev[a]

[a] Department of General Chemistry, Belarusian State Medical University, Minsk, Dzerzinskogo, 83, Belarus
[b] Biochemical Group of the Multidisciplinary Diagnostic Laboratory, Institute of Physiology of the National Academy of Sciences of Belarus, Minsk, Academicheskaya, 28, Belarus

ARTICLE INFO

Article history:
Received 4 November 2017
Received in revised form 21 April 2018
Accepted 23 April 2018
Available online 6 May 2018

Keywords:
Protein
Secondary structure
Propensity scale
Stability
Computer algorithm

ABSTRACT

The influence of flanking structures (alpha helices and beta strands in the primary sequence) on amino acid content of the elements of secondary structure has been analyzed in seven sets of nonhomologous proteins. Elevated usage of beta structural amino acid residues and pentapeptides in beta strands between two alpha helices can be explained by the stabilization of secondary structure of those beta strands by natural selection. High usage of alpha helical amino acids and pentapeptides in beta strands situated between two other beta strands is an evidence of the relaxation of natural selection: "passive" beta strands in these fragments of polypeptide chains are frequently formed due to the influence of flanking "active" beta strands. Alpha helices situated between alpha helix and beta strand are enriched by alpha helical pentapeptides and have lower usage of beta structural pentapeptides than those situated between beta strand and alpha helix, their N-termini are more frequently protected from helix to beta transitions by Glu residues. Alpha helices between two beta strands are stabilized by Ala residues and pentapeptides specific to alpha helices. Dipeptide content of the most stable alpha helices and beta strands was used for the creation of the PentaFOLD 2.0 algorithm (http://chemres.bsmu.by) that finds stable fragments of these elements of secondary structure in PDB files.

1. Introduction

Current methods of secondary structure prediction are thought to perform with the maximal possible accuracy (Crooks and Brenner, 2004). Propensity-based methods (Combet et al., 2000) are working with approximately 70% of efficiency, while homology-based methods show the efficiency above 80% (Drozdetskiy et al., 2015). Two approaches to the prediction of secondary structure may be combined in the same algorithm that uses propensity scales only in a lack of a good template (Sakthivel and Habeeb, 2015; Kandoi et al., 2017). Computer algorithms for 3D modeling of peptides and proteins use the output of algorithms for secondary structure prediction (Xu and Zhang, 2012; Shen et al., 2014). So, during the modeling they postulate that tertiary structure is determined by secondary structure, while tertiary structure may also influence the formation of secondary structure (Zhang and Su, 2012). Some elements of secondary structure are formed both in a full-length protein, and in the short separate peptide. Such fragments have high intrinsic propensity to form certain secondary structure. In contrast, the formation of some other secondary structure elements depends only on long distance interactions. Such elements of secondary structure are quite unpredictable for propensity-based methods, while homology-based ones can still find similar structures with the same kind of long distance interactions (Zhang, 2009). In this study we decided to apply the knowledge on secondary structure formation to the creation of a method that can find the most stable (most predictable) fragments of proteins. To create this method one needs to select the most predictable types of alpha helices, beta strands and random coil regions.

Previously we showed that amino acid content of random coil regions demonstrates strong dependence on their flanking structures in the primary sequence (Khrustalev et al., 2013). We

* Corresponding author at: Belarusian State Medical University, Dzerzinskogo 83, Minsk, Belarus.
E-mail address: vvkhrustalev@mail.ru (V.V. Khrustalev).

https://doi.org/10.1016/j.biosystems.2018.04.002
0303-2647/

classified those regions in four types: coil between two alpha helices (HCH), coil between two beta strands (BCB), coil between alpha helix and beta strand (HCB), and coil between beta strand and alpha helix (BCH). In the current study we have classified alpha helices and beta strands in the same manner: HHH, BHB, HHB, and BHH for alpha helices; HBH, BBB, HBB, and BBH for beta strands. This classification is based on positions of alpha helices and beta strands in the primary sequence, and not in 3D motifs.

The aim of this work is to find the most stable types of alpha helices and beta strands, and to use the features of these elements of secondary structure for the creation of computer algorithm.

The term "stability" in this work does not have a pure thermodynamic sense (Schymkowitz et al., 2005; Hilser and Whitten, 2014), but a probabilistic one. The more stable alpha helix (or beta strand) in this work means the most predictable one with the help of propensity scales based on specific combinations of definite amino acid residues (Combet et al., 2000) and clusters of hydrophobic and hydrophilic amino acids (Lim, 1974; Khrustalev and Barkovsky, 2012).

Indeed, each amino acid residue shows more or less sharp preference to be included in a certain type of secondary structure element, as it was confirmed in numerous studies (Combet et al., 2000; Sakthivel and Habeeb, 2015; Lim, 1974; Chou and Fasman, 1978). In the same manner, combinations of hydrophobic and hydrophilic residues in pentapeptides usually show clear preferences either for alpha helices, or for beta strands, or for random coil (Khrustalev and Barkovsky, 2012). The central idea of the current work is as follows: one can distinguish intrinsic alpha helices, beta strands and regions of coil from those formed because of the influence of other parts of the protein. Intrinsic elements of secondary structure have specific amino acid combinations that make them stable. Such intrinsic alpha helices should be formed on their own, and such intrinsic beta strands should form beta structure on their own together with partners that may not even have specific amino acid content.

2. Materials and methods

This work is based on the analysis of 3D structures of 1513 proteins. There are seven groups of these proteins: 218 Mn^{2+} binding proteins from both bacteria and eukaryotes (Khrustalev et al., 2016a); 353 Mg^{2+} binding proteins from both bacteria and eukaryotes (Khrustalev et al., 2016a); 542 bacterial proteins (Khrustalev et al., 2014); 100 eukaryotic alpha helical proteins (Poboinev et al., 2017); 100 eukaryotic beta structural proteins (Poboinev et al., 2017); 100 eukaryotic "alpha + beta" proteins (Poboinev et al., 2017); 100 eukaryotic "alpha/beta" proteins (Poboinev et al., 2017). Each group contains only nonhomologous proteins. Initial sets of proteins had been larger, but we removed similar proteins. With the help of "Decrease Redundancy" algorithm (https://web.expasy.org/decrease_redundancy) we selected just those proteins that have the percent of similarity between their sequences lower than 25%. Three first groups include proteins from bacteria with low, average and high GC-content of their genomes. So, we used GC-content of the bacterial genome as a criterion to include a protein into the data set: this parameter is relatively constant along the length of the whole bacterial genome. We were unable to determine the structural class for about a half of the bacterial proteins, since this information has not been included in the description from Protein Data Bank. Bacterial proteins with known structural class include mostly "alpha/beta" and "alpha + beta" proteins (Khrustalev et al., 2016a; Khrustalev et al., 2014). To select eukaryotic proteins, we used structural class as a criterion to include them into data sets. This time we were unable to use GC-content of genes as a second criterion, since the amount of G + C usually varies greatly along the same eukaryotic mRNA.

The boarders of the elements of secondary structure have been determined by the DSSP (Dictionary of the secondary structure of proteins) method (Kabsch and Sander, 1983). DSSP is the most widely used method to assign secondary structure for 3D structures of proteins obtained by X-ray and NMR analysis. DSSP does not predict secondary structure, but finds hydrogen bonds between main chain atoms and estimates their patterns (Kabsch and Sander, 1983). The data on secondary structure have been converted to the 3-type secondary structure description: we distinguish alpha helices, beta strands, and everything else as random coil. The next step was to classify beta strands into four types according to the flanking elements of secondary structure in the primary sequence (and not in 3D motifs): beta strands between two other beta strands (BBB); beta strands between two alpha helices (HBH); beta strands between beta strand and alpha helix (BBH); beta strands between alpha helix and beta strand (HBB). In the same manner we classified alpha helices into four types according to the positions of alpha helices and beta strands in the primary sequence: alpha helices between two beta strands (BHB); alpha helices between two alpha helices (HHH); alpha helices between alpha helix and beta strand (HHB); alpha helices between beta strand and alpha helix (BHH). Random coil regions have also been classified into four types: random coil between two beta strands (BCB); between two alpha helices (HCH); between alpha helix and beta strand (HCB); between beta strand and alpha helix (BCH). According to the structural classification of proteins, alpha helical proteins may still have a few beta strands, while beta structural proteins may still have a few alpha helices. All the types of alpha helices, beta strands and random coil regions can be found in all the seven groups of proteins, while, for example, in beta structural proteins the number of HHH regions is lower than the number of BHH or HHB regions, and much lower than the number of BHB regions.

In this study we focused our attention on the effects of co-translational folding. So, we considered the order of secondary structure elements in the primary sequence only, and not in 3D motifs. For example, if we use the term "HHH", we mean that it is an alpha helix that is flanked by another alpha helix in the direction towards N-terminus and by another alpha helix in the direction towards C-terminus. An alpha helix from "HHH" motif may interact with two flanking helices, it may interact with just one of them, or it may not interact with them at all. In the same manner, a beta strand from "BBB" motif may not form beta structure with two flanking beta strands, it may form it with just one of them, or it may be a part of a beta sheet made from three or more strands.

The data on the length of random coil regions between alpha helices and beta strands in seven sets of proteins can be found in the Supplementary Material file "Length of random coil.xlsx". From 77.82–93.55% of random coil regions have a length from 1 to 10 amino acid residues. Long regions of random coil (longer than 20 amino acid residues) are rare: they contribute from 1.01 to 6.45% into our data sets. We did not include in our study those regions of random coil that have amino acid residues with undetermined locations. This technique helped to avoid the usage of intrinsically disordered regions that are usually rather long. We did not use any minimal or maximal values for the length of random coil between alpha helices and beta strands to classify the motifs, since the main idea of the study was to check how the next and the previous element of secondary structure influence the amino acid content of alpha helices and beta strands situated between them in the primary sequence, and not to study linear fragments of polypeptide chain that form certain types of 3D motifs.

We calculated amino acid content in each type of beta strands, alpha helices, and random coil regions, and compared the usage