Biochemical and Biophysical Research Communications 472 (2016) 209–216

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

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

Chameleon sequences in neurodegenerative diseases

Golnaz Bahramali [a], Bahram Goliaei [a, *], Zarrin Minuchehr [b, **], Ali Salari [b]

[a] Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran
[b] Department of Systems Biotechnology, National Institute of Genetic Engineering and Biotechnology, (NIGEB), Tehran, Iran

ARTICLE INFO

Article history:
Received 23 January 2016
Accepted 30 January 2016
Available online 23 February 2016

Keywords:
Protein secondary structure
Chameleon sequences
Neurodegenerative diseases
Sequence properties
Enrichment analysis

ABSTRACT

Chameleon sequences can adopt either alpha helix sheet or a coil conformation. Defining chameleon sequences in PDB (Protein Data Bank) may yield to an insight on defining peptides and proteins responsible in neurodegeneration. In this research, we benefitted from the large PDB and performed a sequence analysis on Chameleons, where we developed an algorithm to extract peptide segments with identical sequences, but different structures. In order to find new chameleon sequences, we extracted a set of 8315 non-redundant protein sequences from the PDB with an identity less than 25%. Our data was classified to “helix to strand (HE)”, “helix to coil (HC)” and “strand to coil (CE)” alterations. We also analyzed the occurrence of singlet and doublet amino acids and the solvent accessibility in the chameleon sequences; we then sorted out the proteins with the most number of chameleon sequences and named them Chameleon Flexible Proteins (CFPs) in our dataset. Our data revealed that Gly, Val, Ile, Tyr and Phe, are the major amino acids in Chameleons. We also found that there are proteins such as Insulin Degrading Enzyme IDE and GTP-binding nuclear protein Ran (RAN) with the most number of chameleons (640 and 405 respectively). These proteins have known roles in neurodegenerative diseases. Therefore it can be inferred that other CFP's can serve as key proteins in neurodegeneration, and a study on them can shed light on curing and preventing neurodegenerative diseases.

1. Introduction

Anfinsen proposed that the proteins are predisposed to fold into a unique three dimensional structure which is clearly specified by its amino acid sequence, according to Anfinsen's dogma, the amino acid sequence of a protein contains sufficient information to determine its three-dimensional structure [1]. This broadly accepted theory was used as the central dogma in predicting the secondary and tertiary structures from its sequence alone following the pioneer work of Chou and Fasman [2]. Anfinsen's theory was shaken by the discovery of the chameleon sequences which can fold as different secondary structures in proteins, these sequences are abundant in nature and play a crucial role in human diseases and are said to constitute as part of the proteome named ‘unfoldome’ [3]. These segments with ambivalent structures were first reported by Kabsch and Sander [4]. Examples of related studies are the involvement of chameleon sequences in the induction of misfolding diseases such as amyloid fibril formation of neurodegeneration [5–8]. Neurodegenerative diseases including Alzheimer's, Parkinson's, Huntington's, Creutzfeldt-Jakob disease, etc. involve a series of brain proteins named amyloid proteins, which upon interaction of the neuronal membranes monomers of amyloid proteins undergo an alpha helix to sheet shift in their conformation. These sequences are also suggested to be one of the limiting factors for the accuracy of secondary structure prediction methods and one important reason for misprediction of programs designed for protein secondary structures is the structural diversity among the peptides with the same sequence i.e. the chameleons [9–13]. The relationship of an amino acid sequence to its eventual structures is important for the structural prediction and design purposes as well as for the comprehension of diseases caused by a protein conformation [10,14,15], and the identification of the propensity values would provide local sequence information for predicting secondary structures [16,17].

Many studies showed that the sequence neighboring in secondary structure of proteins is important in forming these particular structures [18,19]. It has been mentioned that the amino acid propensities for secondary structures can be still improved to obtain better predictive results and to reveal important structural information's [20]. Different amino acids have different preferences

[*] Corresponding author.
[**] Corresponding author.
E-mail addresses: goliaei@ut.ac.ir (B. Goliaei), minuchehr@nigeb.ac.ir (Z. Minuchehr).

http://dx.doi.org/10.1016/j.bbrc.2016.01.187
0006-291X/© 2016 Elsevier Inc. All rights reserved.

for their neighbors and that these local interactions are crucial for their structural conformation, and are also used in the secondary structure prediction methods [21–25]. Due to the importance of the chameleon sequences [5,11,13] and the involvement of the local amino acid interactions in secondary structure formation, we hereby present a comprehensive meta-analysis of single and double propensity of amino acids in chameleon sequences of the Protein databank (PDB), in order to find proteins with the most number of chameleons and name them Chameleon Flexible Proteins (CFPs). We have also built different chameleon groups corresponding to helix, strand and coil, along with the analysis of their solvent accessibility, presenting their singlet and doublet amino acid propensities.

2. Methods

2.1. Databases

A 8315 non-redundant protein chains in the PDB database was used for gathering the chameleon sequences. This set was generated from the current version of the PDB (Dec. 2014) [26] using the PISCES protein sequence algorithm [27] which provided the most up-to-date collection with the following criteria of non-redundant PDB chain database by the selection method of Hobohm et al. [28].

Experimental method = X-ray crystallography, maximum resolution 2.5 A°, maximum R-value 0.3, maximum sequence percentage identity = 25% or less. To avoid statistical bias caused by the large number of homologues proteins this dataset was used for our subsequent statistical analysis.

2.2. Chameleon sequence determination

Secondary structure assignments were made automatically using the DSSP [4] program. The 8 level secondary structural assignments in DSSP were reduced to the 3 classical states: helix including α, 3_{10} and π-helices, strand the β -strand assignments, and coil which covered the rest of our assignments (γ-bridges, turns, bends and coils). For both datasets, three non-redundant sequence files were prepared based on DSSP ($\geq$4 amino acids). Our first file was sequences with helix conformations only, our second file was β-strands and the third was restricted to the coils or unstructured sequences. The tool for extracting segments with identical sequences and complete different secondary structures was designed using our in house C-sharp program. In this process, we found the helix sequences (list H) with the strand (list E) and the coil sequences (list C) by sliding the helix sequence along the strand sequence, one residue at a time. In addition, sequences which corresponded to the entire strands were searched against the helix and the coil sequences, and sequences that correlated with the entire coils were searched against the helix and strand sequences. Finding the same sequences was performed as followed: first, we searched for all possible identical 4 residues (4-mer) in one list (e.g. H list) and another list (e.g. E list) using a matching matrix, wherever possible. These residues were then extended to identify longer identical sequence pairs (4–12-mer). Contiguous, overlapping 4-mers that could form higher order n-mers were not retained in the 4-mer dataset and were assigned as the appropriate n-mer while only the longest possible n-mer was considered. Where one sequence from one list exactly matched the target sequence from another list, it was designated as a “chameleon” sequence for the corresponding protein, identified by its PDB code (e.g. HHHHH in one protein and EEEEE in another protein) and then they were classified into 3 distinct groups namely, Helix-Strand (HE-Chameleons), Helix-Coil (HC-Chameleons) and Coil–Strand (CE-Chameleons). For each chameleon peptide, the peptide sequence length, the peptide sequence, the PDB code/chain, the protein name and the location of the chameleon sequence along the protein chain were recorded. Finally we sorted the proteins for their number of chameleons, in order to find the most flexible proteins in the protein databank.

2.3. Residue occurrence

In order to avoid biases in the statistical analyses, the following survey was accomplished on the chameleon sequences in our dataset (sequence identity less than 25%). To investigate the residue occurrences in the extracted chameleon peptides dataset, the amino acid frequencies were calculated from all n-mers in dataset that had undergone complete helix to strand (HE), helix to coil (HC) and strand to coil (CE) transitions. These values were normalized against the occurrence of the amino acid frequencies in the two types of structure involved in our dataset.

In order to calculate the amino acid neighboring preferences, we used the following methodology; for 20 amino acids, there were 400 possible amino acid doublets (i.e., neighbors). For amino acid i, all 20 nij values (with j = 1, 2,. . . 20, corresponding to the 20 amino acids), provided a profile of neighbor preference for amino acids found after amino acid i while all 20 nij values provided another profile for amino acids found before amino acid i along the amino acid sequence. Additionally, the situation of every doublet was analyzed. We assumed the neighbor-dependent propensity values as Σx $(a \pm 1)$ where the Σx $(a \pm 1)$ value of 1.0 means that the occurrence of the residue pair, a*x* (or *x*a), in the chameleon sequences is the same as its frequency of occurrence of the amino acid neighboring in the two types of structure involved in database. A value > 1.0 means that the pair has an occurrence in the chameleon sequences which is higher than its incident in the PDB, suggesting that the pair has a preference for adopting chameleon sequences. Furthermore, Σx $(a \pm 1)$ values lower than unity suggest less preference for the pair in the chameleon sequences, all of our singlet and doublet propensity calculations were mentioned in our previous studies [25,29].

2.4. Solvent accessibility analysis

The solvent accessibility of each segment is the solvent accessibility value per residue as computed by the DSSP program averaged over the segment's length. The relative solvent accessibility of each residue was estimated by normalizing the absolute value by the maximum accessibility per residue. In this work, we assigned two values (i.e. buried (B) and exposed (E)), depending on the average accessibility value, for either being higher (or equal) and lower than the 16% threshold, respectively [30,31].

2.5. Enrichment analysis of chameleon sequences

In order to investigate the human disease enrichment analysis of HE-, CE- and HC-Chameleon sequences were performed using interactive and collaborative gene list enrichment analysis tool (Enrichr: http://amp.pharm.mssm.edu/Enrichr/) [32]. Enrichr is an integrative web-based software application that includes 35 gene-set libraries, an approach to rank enriched terms. To find the disease categories, three gene set library databases such as OMIM (Online Mendelian Inheritance in man) [33], Disease Perturbations from GEO (Gene Expression Omnibus) up and Disease Perturbations from GEO down were used. In the results section of this tool, the computed p-value was combined using the Fisher exact test with the z-score of the deviation from the expected rank and produced a combined score rank. To gain insight into the potential map pathway of the proteins with chameleon sequences, KEGG