Biochimica et Biophysica Acta 1824 (2012) 637–646

Contents lists available at SciVerse ScienceDirect

Biochimica et Biophysica Acta

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

Charged single alpha-helices in proteomes revealed by a consensus prediction approach

Zoltán Gáspári [a,1], Dániel Süveges [b,2], András Perczel [a,c], László Nyitray [b,*], Gábor Tóth [d,**]

[a] Institute of Chemistry, Eötvös Loránd University, Pázmány Péter sétány 1/A, 1117 Budapest, Hungary
[b] Department of Biochemistry, Eötvös Loránd University, Pázmány Péter sétány 1/C, 1117 Budapest, Hungary
[c] ELTE-HAS Protein Modelling Group, Pázmány Péter sétány 1/A, 1117 Budapest, Hungary
[d] Agricultural Biotechnology Center, Szent-Györgyi Albert u. 4, 2100 Gödöllő, Hungary

ARTICLE INFO

Article history:
Received 2 October 2011
Received in revised form 20 January 2012
Accepted 22 January 2012
Available online 28 January 2012

Keywords:
Charged single alpha-helix
Structure prediction
Protein evolution

ABSTRACT

Charged single α-helices (CSAHs) constitute a recently recognized protein structural motif. Its presence and role is characterized in only a few proteins. To explore its general features, a comprehensive study is necessary. We have set up a consensus prediction method available as a web service (at http://csahserver.chem.elte.hu) and downloadable scripts capable of predicting CSAHs from protein sequences. Using our method, we have performed a comprehensive search on the UniProt database. We found that the motif is very rare but seems abundant in proteins involved in symbiosis and RNA binding/processing. Although there are related proteins with CSAH segments, the motif shows no deep conservation in protein families. We conclude that CSAH-containing proteins, although rare, are involved in many key biological processes. Their conservation pattern and prevalence in symbiosis-associated proteins suggest that they might be subjects of relatively rapid molecular evolution and thus can contribute to the emergence of novel functions.

1. Introduction

The charged single α-helix (CSAH)[3] is a recently identified universal protein structural motif [1–3]. Such helices, formed by sequences harboring a high fraction of charged residues with a characteristic alteration of positive and negative charges, are stable in their monomeric form in aqueous solution, unlike other helical fragments of proteins. It has been proposed that CSAHs are stabilized by the interplay between short- and long-range electrostatic interactions [4–6].

CSAHs have been shown to be present in a wide range of proteins and were suggested to play diverse roles such as mediating transient interactions and/or acting as relatively rigid spacers or extensions [2]. The role of the CSAH region in myosins VI and X as an extension of the lever arm has been studied extensively [5–8].

CSAH-forming sequences exhibit low complexity, are rich in repetitive residue segments and have a high fraction of charged residues, most prominently Glu, Arg and Lys [1,2]. CSAHs are therefore often predicted either as intrinsically unstructured segments, as coiled-coils or both. These cross-predictions might indicate evolutionary transitions between these motifs and can also have functional relevance in terms of the plasticity and capability for structural rearrangements of these segments [9]. Nevertheless, as CSAHs constitute a structural motif distinct from the above mentioned ones with their stable monomeric helical form, we previously developed dedicated detection methods for their prediction from protein sequences [2]. In this study we set up a consensus prediction interface (available as a web server and as a set of standalone tools) incorporating our conceptually unrelated methods (SCAN4CSAH and FT_CHARGE), and perform a detailed analysis of sequences deposited in the UniProt database [10].

2. Methods

2.1. Consensus CSAH detection

The conceptual basis of both SCAN4CSAH and FT_CHARGE has been described earlier [2]. In brief, SCAN4CSAH applies a scoring scheme based on the expected stabilizing/destabilizing effect of different patterns of charged side chains of specific sequential distances. Dyads of oppositely charged residues three or four positions apart and triads of alternately charged side chains at

* Correspondence to: L. Nyitray, Department of Biochemistry, Eötvös Loránd University, Pázmány Péter sétány 1/C, 1117 Budapest, Hungary. Tel.: +36 1 2090555x8783; fax: +36 1 3812172.
** Correspondence to: G. Tóth, Agricultural Biotechnology Center, Szent-Györgyi Albert u. 4, 2100 Gödöllő, Hungary. Tel.: +36 28 526224; fax: +36 28 526101.
E-mail addresses: gaspari.zoltan@itk.ppke.hu (Z. Gáspári), daniel.suveges@ucsf.edu (D. Süveges), perczel@chem.elte.hu (A. Perczel), nyitray@elte.hu (L. Nyitray), tothg@abc.hu (G. Tóth).
[1] Present address: Faculty of Information Technology, Pázmány Péter Catholic University, Práter u. 50/A, 1083 Budapest, Hungary.
[2] Present address: Department of Cellular and Molecular Pharmacology, University of California, San Francisco 600 16th Street, MC 2140, San Francisco, CA 94158-2140, USA.
[3] Abbreviations used in the text: CSAH: charged single α-helix and GO: gene ontology.

1570-9639/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.bbapap.2012.01.012

Table 1
Protein sequence data sets used in this study.

Name	Description
SwissProt70	SwissProt sequences filtered at 70%
SwissProtCSAH70	All CSAH-containing sequences in SwissProt filtered at 70%
UniProtCSAH70	All CSAH-containing sequences in UniProt filtered at 70%
UniProt_HUMAN_70	The human proteome set filtered at 70%
UniProt_MOUSE_70	The mouse proteome set filtered at 70%
UniProtCSAHPerorganism70	CSAH-containing sequences were extracted for each organism and filtered at 70% independently of sequences from other organisms

positions $i-i+4-i+8$, $i-i+3-i+7$ or $i-i+4-i+7$ are regarded as stabilizing. In contrast, identically charged residues placed $i-i+3$, $i-i+4$ and oppositely charged ones at $i-i+1$, $i-i+2$ positions are taken into account as destabilizing interactions. Scoring of these patterns was optimized to favor several selected known CSAH segments. The scores were ultimately turned to probabilities (P-values) by fitting an extreme value distribution (EVD) [11,12] to the data [2].

The FT_CHARGE method detects CSAHs by analyzing the amplitudes and frequencies in the Fourier transform of the charge correlation function calculated for a given segment [2]. CSAHs typically have frequencies between 1/6 and 1/9. For the current improved version of FT_CHARGE, random segments containing Ala, Arg and Glu only were generated with lengths of 16–128 residues according to the powers of 2 and with different compositions by changing Arg and Glu content by 10% at each step. For each of these parameters, 5000 sequences were generated and evaluated, and an EVD was fitted to the resulting maximum amplitudes (Supplemental Figure S1). These distributions allow the score of a submitted segment to be converted to a P-value relevant for its length and composition.

We have defined the consensus of the two methods as segments identified by both methods with a minimum length that corresponds to the lower threshold set for the methods. For example, SCAN4CSAH uses a default minimum length of 40, whereas FT_CHARGE employs windows of 32 and 64 residues by default and combines the results obtained with these. Thus, an overlap of the predictions over at least 32 residues is required for the consensus-based identification of a CSAH. FT_CHARGE applies a sliding window approach with a sliding parameter set to 1 as default as this ensures precise definition of CSAH boundaries.

2.2. Analysis of CSAH distribution in UniProt (version 2011_05)

Since FT_CHARGE is computationally more demanding than SCAN4CSAH, CSAH detection on large data sets is done in two steps: first, potential CSAH-bearing sequences are identified by SCAN4CSAH, and FT_CHARGE is run only on these with window sizes 32 and 64 and a sliding parameter of 1. Overlapping FT_CHARGE-detected segments are then combined regardless of the window size they were identified with. Finally, CSAHs matching the consensus length criterion (see above) are extracted using the outputs of both detection algorithms. This approach is implemented in the 'csahdetect.pl' script downloadable from the CSAH server web site.

In order to standardize the methods for analysis, we created a UniProt-style database for CSAH-containing sequences by simply adding CSAH annotation lines to all relevant entries. These files are provided as a CSAH database (CSAHdb) at the csahserver web site (http://csahserver.chem.elte.hu).

To analyze the overlap between CSAH/coiled-coil and CSAH/disorder predictions, we used the standalone versions of the COILS (available as 'ncoils', [13] and the IUPred [14,15] programs with default parameters. In functional analyses, we considered the domain annotation along with GO term listing provided in SwissProt. Expected number of proteins with co-occurrence of two domain types was calculated by assuming total independence of the domains as $N_{d1} \times N_{d2} / N_{all}$, where N_{d1} and N_{d2} are the number of proteins containing domains d1 and d2, respectively and N_{all} is the total number of proteins in the dataset. Expected number of proteins with a given GO term was calculated in an analogous way. To obtain a P-value of describing the association of domains or GO terms with the presence of CSAHs, we applied the Fisher's exact test as described in ref. [16] using the R statistics software [12].

Where appropriate, sequence sets were filtered at 70% similarity using the CD-HIT [17] program. Depending on the nature of the analysis, either the filtered version of full SwissProt database or filtered (sub)sets of CSAH-containing sequences were used (Table 1).

3. Results and discussion

3.1. Description of the CSAH server

Original description of the detection algorithms can be found in ref. [2], and more recent improvements are detailed in Section 2.1. The CSAH server incorporates both SCAN4CSAH and FT_CHARGE

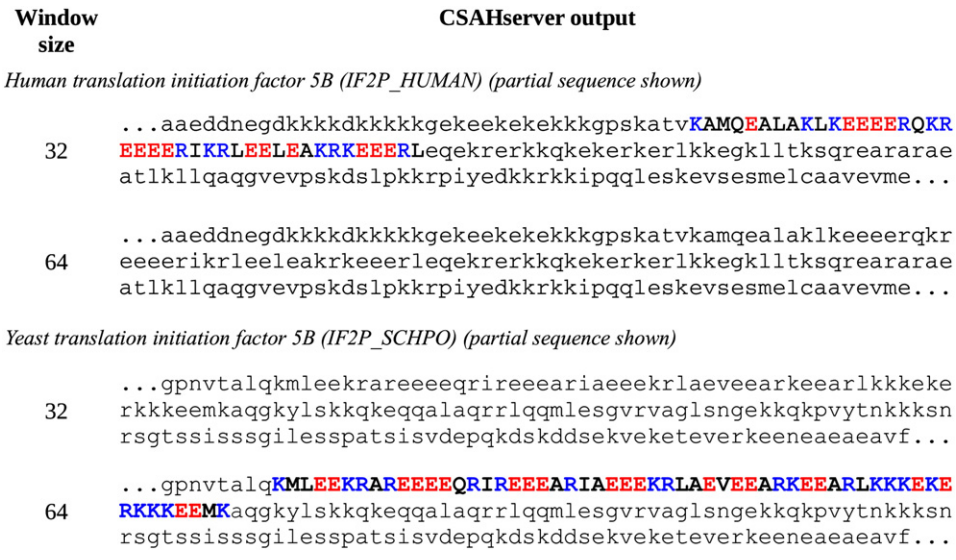

Fig. 1. Examples of CSAH server outputs where CSAH detection depends on the window size applied for FT_CHARGE. The consensus identified CSAHs are shown (SCAN4CSAH was run with default parameters, FT_CHARGE with step size 1 and window sizes as specified).