Preprint of: Biochem. Biophys. Acta - Proteins and Proteomics (2012) 1824:637-646. [http://dx.doi.org/10.1016/j.bbapap.2012.01.012]

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,*] and 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

[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

Corresponding authors:
László Nyitray
Department of Biochemistry, Eötvös Loránd University, Pázmány Péter sétány 1/C, 1117 Budapest, Hungary
Phone: +36-1-2090555 ext. 8783
Fax: +36-1-3812172
E-mail: nyitray@elte.hu

Gábor Tóth
Agricultural Biotechnology Center, Szent-Györgyi Albert u. 4, 2100 Gödöllő, Hungary
Phone: +36-28-526224
Fax: +36-28-526101
E-mail: tothg@abc.hu

E-mail addresses of all authors:
Z.G.: gaspari.zoltan@itk.ppke.hu
D.S.: daniel.suveges@ucsf.edu
P.A.: perczel@chem.elte.hu
L.N.: nyitray@elte.hu
G.T.: tothg@abc.hu

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.

Keywords: Charged Single Alpha-Helix; structure prediction; protein evolution;

1. INTRODUCTION

The charged single α-helix (CSAH)[1] 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 [4,5,6].

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 [5,6,7,8].

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].

1Abbreviations used in the text: CSAH: charged single α-helix, GO: gene ontology

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 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 1/6-1/9. 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 (Supplemetal 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 $N_{d1}*N_{d2}/N_{all.}$, 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 subsection 2.1. The CSAH server incorporates both SCAN4CSAH and FT_CHARGE and is available at *http://csahserver.chem.elte.hu*. The server accepts a single protein sequence as input, either pasted into the input field, uploaded as a FASTA file, or specified by a UniProt ID. The server reports the consensus of the two methods by highlighting the CSAH regions on the sequence and also in a tabular format. Separate outputs of both SCAN4CSAH and FT_CHARGE are also provided as well as a FASTA format file in which the CSAH regions are masked. Moreover, the user is allowed to select only one of the methods to be performed and several parameters can be customized for both methods. The default parameters chosen for optimum CSAH detection for the SCAN4CSAH method are: minimum CSAH segment length of 40 residues with a maximum allowed uncharged gap size of 5. The FT_CHARGE method is invoked by default with a frequency range of 1/9-1/6, two window sizes, 32 and 64 and a step size of 1. The P-value threshold is 0.01 for both programs and is not adjustable from the current web interface. The consensus CSAH cannot be shorter than the lower of the minimum threshold length set for each detection method.

The server allows only a single sequence as input, thus both the SCAN4CSAH and FT_CHARGE methods are provided as downloadable Perl programs along with their Extreme Value Distribution (EVD) parameter files and a wrapper script to make large-scale analysis of sequences possible. In addition, a database of CSAHs (CSAHdb) is available based on SwissProt release 2011_05.

3.2. Effect of FT_CHARGE window length on CSAH identification

Due to the underlying principle of FT_CHARGE, this method exhibits nonlinear dependence on the window size used, meaning that CSAHs identified with shorter window sizes might not necessarily score over the threshold when analyzed with a longer window. This, although surprising at first sight, is a consequence of the FT method being sensitive for periodicities within the entire length of the window examined. In other words, FT_CHARGE tests the segments globally, i.e. assesses its tendency to form a CSAH matching its entire length. CSAHs identified with longer window sizes either exhibit relatively low charge density with high regularity or low regularity with high charge density, whereas those identified only with shorter window lengths do not maintain the characteristic periodicity over longer stretches (**Figure 1**).

The choice of window length also influences whether long CSAHs are recognized as one long segment or as multiple shorter ones, as parts not identified with one window size can interrupt long CSAH candidates detected with the other. To overcome such difficulties, FT_CHARGE can be invoked with window sizes of both 32 and 64 (in a single run as allowed in the present implementation) and all overlapping segments are combined. In the current implementation, this is the default setting.

3.3. Evaluation of the consensus detection

To analyze and demonstrate the correspondence of the two predictors, we have used a homology-filtered set (designated UniProt_human_70, see subsection 2.2.) of the human proteome derived from the collection supplied with UniProt 2011_05. It is clear that SCAN4CSAH is by far more permissive than FT_CHARGE, the latter confirming CSAHs identified by the former (i.e. having an overlap >50%) in only around 3% of the cases (**Figure 2**). In contrast, almost 80% of CSAHs found by FT_CHARGE overlap more than 50% with SCAN4CSAH predictions. It must be noted that

SCAN4CSAH identifies around twenty times more CSAHs than FT_CHARGE in this dataset. In summary, FT_CHARGE is both the more computationally demanding and more restrictive of the two algorithms, which heavily reduces the number of CSAHs identified by the consensus approach.

The consensus detection, i.e. the use of the common part of the regions found by two conceptually unrelated prediction algorithms is expected to minimize the number of false positives, i.e. segments predicted to form CSAHs which actually do not form this type of structure with high probability. On the other hand, this also means that a higher fraction of genuine CSAHs might remain undetected. However, in the absence of extensive experimental information about CSAHs we prefer to present a highly conservative estimate of such elements to avoid overestimation either of their abundance or significance. It should also be emphasized that while typical predictors usually yield a two-state output using cutoff values, protein structural elements including CSAHs are dynamic at different time scales [4] and thus might not be best characterized as distinct well-defined states.

3.4. Overlap of CSAHs with predicted coiled-coil and disordered segments

We have previously noted that CSAHs tend to overlap with annotated coiled-coil and/or intrinsically disordered segments in proteins [2]. In a systematic study on cross-predictions between coiled-coil and disorder-recognizing algorithms [8] we have suggested the use of COILS and IUPred for effective yet minimally overlapping identification of such segments. Here we present the analysis of the identified CSAH segments with respect to COILS and IUPred predictions (subsection 2.2.). In general, overlap of CSAHs with predicted coiled-coil segments is remarkably higher than that with predicted disordered regions (**Figure 3**). It should be noted that the majority of overlaps is either 0 or 100%, the latter being much more common, and intermediate values are relatively rare (**Table 2**). These considerations justify

our approach for developing a specific prediction approach for CSAHs as they cannot be unambiguously identified as segments recognized both as coiled-coils and disordered ones.

3.5. Overview of CSAH-containing proteins in UniProt

Using UniProt release 2011_05, we have performed a large-scale analysis of CSAHs identified by the consensus method (Supplemental Figure S2). For human and mouse we have used the complete (human and mouse) proteomes supplied with the release. The number of CSAHs and CSAH-containing proteins for selected organisms is shown in **Table 3**. It is apparent that CSAHs represent a very rare protein structural motif generally found in less than 0.2% of all proteins of an organism. Here it is essential to note again that our consensus approach is tailored to yield a conservative prediction minimizing the number of false positives at the probable cost of leaving a high fraction of undetected CSAHs (false negatives). CSAHs predicted solely with SCAN4CSAH can be regarded as an upper estimate of their abundance yielding ~2% in full UniProt (14505 proteins, ~2.7% of all sequences in SwissProt; 295018 proteins, ~1.9% of all sequences in TrEMBL). Thus, the expected number of CSAHs is presumably higher than predicted by the consensus approach, although the motif is inevitably a rare one that explains its relatively recent recognition [6].

Intriguingly, *Homo sapiens* is one of the organisms with the highest number of CSAH proteins in their genome (**Table 4**). The lists obtained are remarkably similar at different levels of similarity filtering, i.e. when only proteins differing above a given threshold are considered. Therefore we can safely assume that high ranking of these species is not due to multiple copies of the same protein sequence in the database. It should also be noted that potential pathogens and parasites are prevalent in all three

listings shown. Whether this has any biological implications or simply reflects a bias in database content remains elusive at present and is expected to be justified or rejected only after experimental investigation of the putative proteins predicted to contain CSAHs.

Length distribution of the CSAH segments identified is shown in **Figure 4**. Average CSAH length is ~73 residues in SwissProt and ~82 residues in the full UniProt (calculated on the respective filtered sets of CSAH-containing sequences).

The longest CSAHs are predicted for proteins in the unreviewed TrEMBL data set (**Table 5**). In the absence of experimental verification, the significance of these CSAHs cannot be reliably established. It should be noted that the length of an uninterrupted regular α-helix of 1000 residues is about 150 Å, able to provide a scaffold/filament/spacer for a considerable distance relative to the size of a cell, comparable only to the length of the longest known coiled-coils [18] (not counting multimeric cytoskeletal elements).

There is only one protein with a known structure in which a CSAH was detected, the cytoplasmic domain of the methyl-accepting chemotaxis receptor protein from *Thermotoga maritima* (PDB ID 2CH7 [19]). The CSAH region (residues 400-440 on both chains) is detected near the 'tip' (located most distal relative to the membrane) of a four-helical bundle region (Supplemental Figure S3). Analysis of the structure with the SOCKET web server [20] revealed three short coiled-coil regions, one of which (region 444-489) is adjacent to the CSAH. Besides mediating dimerization, the tip region of helical bundle interacts with the histidine autokinase CheA and the coupling protein CheW. Assessing the significance of the detected CSAH would require further studies.

3.6. Functional analysis of CSAH-containing proteins

GO term analysis of CSAH-containing proteins is summarized in **Table 6**. Our data reveal that CSAHs are preferentially found in proteins promoting symbiosis, RNA processing, translation initiation, nucleotide / RNA binding and localized in or around the nucleus in eukaryotes. On the other hand, CSAH motifs are underrepresented in extracellular and membrane proteins. A closer look at the symbiosis-related sequences reveals that all CSAH-containing proteins in this group are from *Plasmodium falciparum* isolate 3D7 (SwissProt organism ID: PLAF7). From the 54 proteins with this GO term in the 70% filtered SwissProt dataset (SwissProt70, see subsection 2.2.), 51 belong to this organism. Still, from the 10 CSAH-containing proteins from this strain, 7 are associated with this GO term, more than expected based on the total number of PLAF7 proteins in SwissProt70 (130).

There are 50 different domains occurring in CSAH-containing proteins in full SwissProt and 45 when considering the 70% filtered set (SwissProt70) only (Supplemental Figure S4). In accordance with the GO term analysis, there are 9 proteins (8 with unique protein name) containing an RRM (RNA recognition motif) domain in SwissProt70 (**Table 7**). When counting only different SwissProt protein names (“mnemonic protein identification codes” according to the SwissProt manual, corresponding to the first part of SwissProt entry names) to exclude bias caused by considering multiple orthologs, the second most abundant domain is the myosin head-like and the protein kinase domain. When considering all proteins that contain the domain types found in CSAH-containing sequences, none of them seems to be associated preferentially with CSAHs based on the analysis of observed/expected ratios (**Table 8**, either the observed/expected ratio or the number of actual proteins with the domain combination is below 3, rendering the unambiguous detection of association unreliable). Interestingly,

despite its abundance in CSAH-containing proteins, the protein kinase domain clearly disfavors CSAH segments.

3.7. Protein variants and modifications affecting CSAHs

To assess whether the length, presence or absence of CSAH segments in proteins is different in various protein isoforms, we analyzed the complete human and mouse proteomes (**Table 9**). It is apparent that in many CSAH-containing proteins the CSAH segment is affected by alternative splicing giving rise to CSAH-less variants. As the presence of CSAH segments might influence e.g. the distance between and the simultaneous binding ability of the two segments at the CSAH termini, these isoforms might well exhibit different and specific functions in different processes.

Effect of modified residues on the CSAH segment was analyzed on the full SwissProt dataset (**Table 10**). Although at first sight it might be surprising that changing a single residue can affect the presence of a CSAH, we note that FT_CHARGE inspects the full charge pattern globally and small changes in it might push an otherwise low maximum observed amplitude below the detection limit. Although it can be reasonably expected that such a segment most likely retains some of its helix-forming propensity and the observed effect is due to the sensitivity of our prediction methods, there are known cases where mutating a single residue causes major fold change in globular proteins [21]. Interestingly, our survey does not indicate serious conformational changes upon phosphorylation in spite of the introduction of a negative charge by this post-translational modification. The presence of dimethylated arginine in the proline- and glutamine-rich splicing factor SFPQ is consistent with the notion that arginine methylation is common in proteins involved in RNA-processing [22].

3.8. CSAHs in related proteins

We have investigated the presence and length of CSAH segments in several selected protein families and related protein pairs. Most notably, many CSAHs are found in both bacterial and eukaryotic translation initiation factors suggesting that CSAHs are common in such proteins. However, only 173 of the 743 bacterial IF2 proteins in SwissProt 2011_05 were found to contain CSAHs by the consensus detection method. Sequence alignment of the CSAH-containing region of selected IF2 proteins reveals a remarkably low level of sequence conservation in these segments (**Figure 5**), indicating that CSAHs are not indispensable parts of IF2 proteins.

Myosins show a similarly low level of sequence conservation in their CSAH region (**Figure 6.**). Moreover, the exact location of CSAH segments is heavily dependent on myosin families. In the myosin X family the detected CSAHs are even outside the region shown in **Figures 6 and 7**. Our consensus method predicts CSAH only in bovine myosin X, but by setting the FT_CHARGE window size to 16, a short CSAH in human myosin X can also be detected.

It is apparent that CSAHs overlap with differently predicted regions in IF2s and myosins: in IF2s they coincide with predicted disordered regions and in myosins with coiled-coils (**Figure 8**). It is also evident that CSAHs do not constitute a highly conserved motif in neither of the families. In several myosins, CSAHs contribute to the specific properties of the lever arm (length and stability), which is variable even among myosin subfamilies with no CSAHs.

In the *E. coli* IF2, the detected CSAH region occupies residues 163-217 which is present in all currently known isoforms and according to circular dichroism studies, exhibits a largely helical conformation [23]. No atomic resolution structure is available for this segment and it was not placed in the cryo-EM reconstruction of the *E. coli* translation initiation complex either [24]. The exact role of

this region is not clear, the broader region has been shown to interact both with the ribosome and *infB* mRNA and regarded as a flexible linker between the N-terminal domain (IF2N) and the highly conserved C-terminal region [25]. We propose that the CSAH segment could act as a relatively rigid spacer between two interaction sites of IF2.

4. CONCLUSIONS

CSAHs constitute a rare yet universal and versatile protein structural motif. They are preferentially found in eukaryotes and in symbionts/parasites. They are most abundant in various RNA-interacting proteins. The motif often coincides with disordered segments and/or coiled-coils and is only weakly conserved in IF2 proteins and myosins. This suggests that CSAH motifs might be rapidly evolving (e.g. by expansion of the underlying DNA-sequence repeats and then by mutation disrupting the charge pattern) and provide a flexible platform in evolutionary sense and provide one extremity of tuning the intrinsic helicity of disordered segments or the dimerization propensity of coiled-coils. Mutations in CSAH segments might promote structural changes such as spacer length between two interaction domains in a largely autonomous way and thus CSAHs might be useful structural elements in pathogenicity-enhancing proteins.

5. ACKNOWLEDGEMENTS

This work was supported by grants from the Hungarian Scientific Research Fund (OTKA F68079, K72973, K61784, NI68466, NK81950), and the International Centre for Genetic Engineering and Biotechnology (CRP/HUN09-03). The European Union and the European Social Fund have provided

financial support to the project under the grant agreement no. TÁMOP 4.2.1./B-09/1/KMR-2010-0003. A János Bolyai Research Fellowship to Z.G. and Ferenc Deák Fellowship to D.S. is also acknowledged.

6. REFERENCES

[1] S. Sivaramakrishnan, B. Spink, A.Y.L. Sim, S. Doniach, J.A. Spudich, Dynamic charge interactions create surprising rigidity in the ER/K α-helical protein motif, Proc. Natl. Acad. Sci. USA 105 (2008) 13356-13361.

[2] D. Süveges, Z. Gáspári, G. Tóth, L. Nyitray, Charged single α-helix: a versatile protein structural motif, Proteins, 74 (2009) 905-916.

[3] M. Peckham, P.J. Knight, When a predicted coiled coil is really a single α-helix, in myosins and other proteins, Soft Matter 5 (2009) 2493-2503.

[4] L.M. Espinoza-Fonseca, D. Süveges, Z. Gáspári, G. Tóth, L. Nyitray, Role of catioinc residues in fine tuning the flexibiliy of charged single α-helices, Biophys J. 96 (2009) 322-322.

[5] J.A. Spudich, S. Sivaramakrishnan, Myosin VI: an innovative motor that challenged the swing lever arm hypothesis. Nat. Rev. Mol. Cell. Biol. 11 (2010) 128-137.

[6] P.J. Knight, K. Thirumurugan, Y. Xu, F. Wang, A.P. Kalverda, W.F. Stafford III, J.R. Sellers, M. Peckham, The predicted coiled-coil domain of myosin 10 forms novel elongated domain that lengthens the head, J. Biol. Chem. 280 (2005) 34702-34708.

[7] B.J. Spink, S. Sivaramakrishnan, J. Lipfert, D. Doniach, J.A. Spudich, Long single a-helical tails bridge the gap between structure and function in myosin VI. Nat. Struct. Mol. Biol. 15 (2008) 591-597.

[8] T.G. Baboolal, T. Sakamoto, E. Forgacs, H.D. White, S.M. Jackson, Y. Takagi, R,E, Farrow, J.E. Molloy, P.J. Knight, J.R. Sellers, M. Peckham, The SAH domain extends the functional length of the myosin lever, Proc. Natl. Acad. Sci. USA 52 (2009) 22193-22198.

[9] B. Szappanos, D. Süveges, A. Perczel, L. Nyitray, Z. Gáspári, Folded-unfolded cross-predictions and protein evolution: the case study of coiled-coils, FEBS Lett. 584 (2010) 1623-1627.

[10] The UniProt Consortium: Ongoing and future developments at the Universal Protein Resource, Nucleic Acids Res. 39 (2011) D214-D219.

[11] A. Stephenson, evd: Extreme value distributions. R News 2, (2002) 31–32.

[12] R Development Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria 2010. [http://www.r-project.org.]

[13] A. Lupas, M. Van Dyke, J. Stock, Predicting coiled coils from protein sequences, Science 252 (1991) 1162–1164.

[14] Z. Dosztányi, V. Csizmók, P. Tompa, I. Simon, The Pairwise Energy Content Estimated from Amino Acid Composition Discriminates between Folded and Intrinsically Unstructured Proteins. J. Mol. Biol. 347 (2005) 827-839.

[15] Z. Dosztányi, V. Csizmók, P. Tompa, I. Simon, IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content**.** Bioinformatics 21 (2005) 3433-3434.

[16] I. Rivals, L. Personnaz, L. Taing, M_C. Potier, Enrichment or depletion of a GO category within a class of genes: which test? Bioinformatics 23 (2007) 401-407.

[17] W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22 (2006) 1658-1659.

[18] Z. Gáspári, L. Nyitray, Coiled coils as possible models of protein structural evolution, Biomol. Concepts 2 (2011) 199-210.

[19] S-Y Park, P.P. Borbat, G. Gonzalez-Bonet, J. Bhatnagar, A.M. Pollard, J.H. Freed, A.M. Bilwes, B.R. Crane, Reconstruction of the chemotaxis receptor-kinase assembly. Nat.Struct. Mol. Biol. 13 (2006) 400-407.

[20] J. Walshaw, D.N. Woolfson, SOCKET: a program for identifying and analysing coiled-coil motifs within protein structures. J. Mol. Biol. 307 (2001) 1427-1450.

[21] P.A. Alexander, Y. He, Y. Chen, J. Orban, PN Bryan. A minimal sequence code for switching protein structure and function, Proc. Natl. Acad. Sci. USA 106 (2009) 21149-21154.

[22] M.T. Bedford, S. Richard, Arginine methylation: an emerging regulator of protein function, Mol. Cell 18 (2005) 263-272.

[23] B.S. Laursen, A.C. Kjærkgaard, K.K. Mortensen, D.W. Hoffman, H.U. Sperling-Petersen, The N-terminal domain (IF2N) of bacterial traslation initiation factor IF2 is connected to the conserved C-terminal domains by a flexible linker, Protein Sci. 13 (2004) 230-239.

[24] G.S. Allen, A. Zavialov, R. Gursky, M. Ehrenberg, J. Frank, The cryo-EM structure of a translation initiation complex from Escherichia coli, Cell 121 (2005) 703-712.

[25] B.S. Laursen, H.P. Sørensen, K.K. Mortensen, H.U. Sperling-Petersen, Initiation of protein synthesis in bacteria. Microbiol Mol Biol Rev 69 (2005) 101-123.

FIGURE LEGENDS

Figure 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).

Figure 2. Overlap of SCAN4CSAH and FT_CHARGE predictions using the human proteome filtered at 70%.

Figure 3. Overlaps of CSAH segments with NCOILS and IUPred predictions. Plots refer to CSAH-containing sequences filtered at 70% similarity.

Figure 4. Lengths of CSAH segments identified with the consensus method. Data compiled from a 70% similarity-filtered set of all CSAH-containing sequences (SwissProtCSAH70 and UniProtCSAH70, subsection 2.2.). The points corresponding to discrete numbers of CSAHs with given length are connected only for better data visualization.

Figure 5. Partial sequence alignment of selected CSAH-containing IF2 proteins with the CSAH region highlighted.

Figure 6. Partial sequence alignment of selected CSAH-containing myosins with the CSAH region highlighted.

Figure 7. Sequence alignment of the CSAH region of bovine and human myosin X with the CSAH region highlighted. The CSAH in myosin X is a consensus of SCAN4CSAH and FT_CHARGE with a window size of 16. Mutations in the human sequence relative to the bovine that affect the charge pattern are underlined.

Figure 8. Overlap of CSAH regions with predicted coiled-coils and disordered segments in IF2 proteins and myosins in the SwissProtCSAH70 dataset.

TABLES

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

Table 2. Mean and median values of the overlap of CSAHs with predicted coiled-coil (ncoils) and disordered (IUPred) segments. Data refer to CSAH-containing sequences filtered at 70% similarity.

	ncoils overlap	**IUPred overlap**
SwissProt		
mean ± standard deviation	77.47 ± 33.98%	60.13 ± 44.85%
median	96.88%	85.94%
inter-quartile range	32.24%	100.00%
Full UniProt		
mean ± standard deviation	73.38 ± 36.55%	66.59 ± 43.05%
median	92.86%	97.73%
inter-quartile range	40.00%	100.00%

Table 3. Number of CSAH segments and CSAH-containing proteins in selected organisms using the consensus CSAH detection method

	all sequences investigated	CSAH segments	sequences with CSAHs	percentage of sequences with CSAHs
All organisms				
SwissProt	528048	586	533	*0.10*
TrEMBL	15062837	8908	8150	*0.05*
All	15590885	9494	8683	*0.06*
Human (*Homo sapiens*)				
SwissProt	20232	50	41	*0.25*
TrEMBL	31106	60	60	*0.19*
All	51338	110	101	*0.20*
Mouse (*Mus musculus*)				
SwissProt	16359	34	32	*0.20*
TrEMBL	29530	70	70	*0.24*
All	45889	104	102	*0.22*
Fruit fly (*Drosophila melanogaster*)				
SwissProt	3120	6	5	*0.16*
TrEMBL	32138	62	51	*0.16*
All	35258	68	56	*0.16*
Thale cress (*Arabidopsis thaliana*)				
SwissProt	10388	7	6	*0.06*
TrEMBL	38766	50	49	*0.13*
All	49104	57	55	*0.11*
Baker's yeast (*Saccharomyces cerevisiae*)				
SwissProt=All	6561	9	9	*0.14*
Escherichia coli K12				
SwissProt	4430	3	3	*0.07*
TrEMBL	75	0	0	*0.00*
All	4505	3	3	*0.07*

Uniprot taxonomic mnemonics used to extract sequences were: human: HUMAN, mouse: MOUSE, fruit fly: DROME, thale cress: ARATH, Baker's yeast: YEAST, *E. coli* K12: ECOLI

Table 4. Organisms with the highest number of CSAH-containing sequences in full UniProt

Unfiltered data set (UniProt)			UniProtCSAHPerOrganism70		
Rank	**CSAH proteins**	**Organism**	**Rank**	**CSAH proteins**	**Organism**
1	**165**	***Homo sapiens* (human)**	1	125	*Plasmodium falciparum (isolate 3D7)*
2	155	*Danio rerio* (zebrafish)	2	110	*Trichomonas vaginalis*
3	138	*Trichomonnas vaginalis*	3	72	*Danio rerio* (zebrafish)
4	138	*Mus musculus* (mouse)	4	59	*Caenorhabditis remanei*
5	127	*Postia placenta strain ATCC 44394 / Madison 698-R)*	5	56	*Toxoplasma gondii*
6	126	*Toxoplasma gondii*	6	**54**	***Homo sapiens* (human)**
7	125	*Plasmodium falciparum (isolate 3D7)*	7	54	*Entamoeba histolytica*
8	90	*Gallus gallus* (chicken)	8	52	*Entamoeba dispar*
9	86	*Rattus norvegicus*	9	51	*Gallus gallus* (chicken)
10	68	*Daphnia pulex*	10	49	*Caenorhabditis briggsae*

The top ten organisms with the highest number of CSAH-containing sequences are shown. Taxon names not corresponding unambiguously to a single species are omitted from the list. Similarity filtering was performed on CSAH-containing sequences of each organism separately.

Table 5. 10 longest CSAHs identified in SwissProt and TrEMBL

	Protein ID	CSAH length	Protein length	Protein description
SwissProt	RB12B_HUMAN	333	1001	RNA-binding protein 12B
	MST1_DROHY	226	344	Axoneme-associated protein mst101(1)
	PERQ2_XENLA	185	1239	PERQ amino acid-rich with GYF domain-containing protein 2
	IF4G_SCHPO	179	1403	Eukaryotic translation initiation factor 4 gamma
	K1211_DANRE	177	1079	Uncharacterized protein KIAA1211 homolog
	GG6L6_HUMAN[1]	170	750	Putative golgin subfamily A member 6-like protein 6
	PERQ2_HUMAN	168	1299	PERQ amino acid-rich with GYF domain-containing protein 2
	CALD1_CHICK	150	771	Caldesmon
	TRHY_HUMAN	146	1943	Trichohyalin
	MNN4_YEAST	143	1178	Protein MNN4

	Protein ID	CSAH length	Protein length	Protein description
TrEMBL	A2FLV9_TRIVA	1769	1775	Viral A-type inclusion protein, putative
	A2F6J1_TRIVA	1683	2624	Putative uncharacterized protein
	Q7QCP0_ANOGA	1175	6668	AGAP002737-PA
	Q4CTJ4_TRYCR	928	1302	Tb-291 membrane-associated protein-like, putative
	D0AAB6_TRYB9	883	1287	Tb-291 membrane associated protein, putative
	A2DUK1_TRIVA	816	1415	Neurofilament protein, putative
	D3B7M2_POLPA	790	1246	Putative uncharacterized protein
	E9ADL3_LEIMA	770	2955	Putative uncharacterized protein
	E3FFP7_STIAD	737	1807	Tetratricopeptide repeat domain protein
	A8X9J5_CAEBR	710	3376	Putative uncharacterized protein

[1] In GG6L6_HUMAN, another CSAH of 153 residues long is detected with the applied parameters, which would be ranked #8 on this list but is omitted to show only the longest CSAH per protein. The two CSAHs in GG6L6_HUMAN are sequentially close but are not predicted to be merged even when applying an FT_CHARGE window length of 16.

Table 6. GO terms associated with CSAH-containing sequences in SwissProt (filtered at 70%)

Expected values are calculated based on the number of all proteins/names in SwissProt and the ratio of CSAH-containing/all proteins. Only terms associated with at least 3 proteins with different SwissProt protein names (mnemonic protein identification codes) are shown.

	All SwissProt70 proteins				Proteins with different name			
GO term	**expected**	**observed**	**obs/exp**	**P-value**	**expected**	**observed**	**obs/exp**	**P-value**
Biological process								
symbiosis, encompassing mutualism through parasitism	0.06	7	115.212.	3.13E-13	0.11	7	62.82	2.20E-11
nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	0.12	4	33.24	7.16E-06	0.1	3	31.11	1.30E-04
cytokinesis	0.2	4	19.79	5.48E-05	0.35	3	8.49	1.00E-02
RNA splicing	0.8	4	5	8.68E-03	0.57	4	6.97	2.70E-03
mRNA processing	1.34	5	3.72	1.15E-02	0.89	5	5.62	2.10E-03
DNA repair	4.33	8	1.85	8.25E-02	1.28	6	4.69	1.90E-03
DNA replication	2.89	3	1.04	7.66E-01	0.83	3	3.63	5.07E-02
translation	13.59	10	0.74	3.99E-01	1.14	4	3.51	2.80E-02
mitosis	1.01	4	3.98	1.87E-02	1.02	3	2.94	8.00E-02
response to stress	1.85	24	12.95	2.20E-016	1.49	3	2.01	1.90E-01
protein phosphorylation	2.33	6	2.57	3.02E-02	3.25	6	1.84	1.50E-01
regulation of transcription	3.95	8	2.03	6.42E-02	4.35	7	1.61	2.10E-01
protein transport	3.01	4	1.33	5.49E-01	1.89	3	1.59	4.40E-01
transcription	6.63	9	1.36	3.19E-01	6.69	8	1.2	5.50E-01
Molecular function								
actin filament binding	0.11	6	54.05	1.83E-09	0.18	5	28.46	1.00E-06
translation initiation factor activity	1.24	37	29.85	2.20E-16	0.22	4	18.48	7.02E-05
motor activity	0.51	3	5.85	1.00E-02	0.33	3	9.15	4.46E-03
rRNA binding	6.53	4	0.61	4.30E-01	0.35	3	8.64	5.22E-03
nucleotide binding	2.04	13	6.36	1.88E-07	1.39	11	7.93	1.77E-07
calmodulin binding	0.41	6	14.46	4.37E-06	0.61	4	6.6	3.33E-03
nucleoside-triphosphatase activity	1.41	5	3.55	1.00E-02	0.71	4	5.62	5.84E-03
actin binding	0.78	9	11.51	1.19E-07	0.9	5	5.54	2.22E-03
RNA binding	7.31	13	1.78	4.00E-02	3.3	12	3.63	1.30E-04
ATP binding	32.95	23	0.7	7.00E-02	12.67	21	1.66	2.52E-02
protein serine/threonine kinase activity	1.8	4	2.22	1.10E-01	2.42	4	1.65	3.11E-01
protein binding	8.7	21	2.42	1.80E-04	14.36	20	1.39	1.24E-01
binding	3.9	5	1.28	4.50E-01	3.82	5	1.31	4.38E-01
zinc ion binding	8.33	11	1.32	2.90E-01	8.54	7	0.82	7.24E-01
DNA binding	13.6	10	0.74	4.00E-01	11.34	9	0.79	6.40E-01
metal ion binding	20.73	3	0.14	2.20E-06	8.1	3	0.37	6.81E-02
Cellular localization								

chromosome, centromeric region	0.17	3	17.71	6.82E-04	0.22	3	13.86	1.38E-03
spindle	0.31	5	16.27	1.60E-05	0.36	3	8.43	5.59E-03
cell cortex	0.22	3	13.8	1.40E-03	0.36	3	8.38	5.68E-03
chromosome	0.81	4	4.93	9.16E-03	0.36	3	8.24	5.97E-03
perinuclear region of cytoplasm	0.49	5	10.16	1.47E-04	0.64	4	6.2	4.14E-03
microtubule	0.83	8	9.6	2.31E-06	0.88	5	5.69	1.98E-03
cell surface	0.58	3	5.17	2.07E-02	0.72	3	4.15	3.63E-02
cytoskeleton	0.63	3	4.77	2.54E-02	0.9	3	3.33	6.21E-02
cell outer membrane	0.92	4	4.33	1.41E-02	0.93	3	3.21	6.78E-02
nucleolus	2.15	5	2.33	6.45E-02	1.76	4	2.28	1.01E-01
cytoplasm	48.38	89	1.84	8.97E-10	15.31	28	1.83	1.56E-03
cytosol	4.3	13	3.03	4.23E-04	7.11	12	1.69	7.79E-02
nucleus	15.53	19	1.22	3.54E-01	18.98	17	0.9	7.11E-01
mitochondrion	3.24	3	0.93	1.00E+00	3.66	3	0.82	1.00E+00
plasma membrane	22.86	12	0.52	1.46E-02	18.34	9	0.49	1.66E-02
integral to membrane	32.24	15	0.47	5.55E-04	30.73	11	0.36	1.33E-05
extracellular region	9.19	4	0.44	8.98E-02	12.24	4	0.33	9.98E-03

Table 7. List of most abundant domains in CSAH-containing proteins

all SwissProt			**SwissProt70**		
Domain name	**No. of proteins**	**No. of proteins with distinct name**	**Domain name**	**No. of proteins**	**No. of proteins with distinct name**
PCI	23	2	RRM	9	8
RRM (RNA Recognition Motif)	17	10	IQ	8	7
IQ	12	8	PCI	8	1
Myosin head-like	10	7	Myosin head-like	6	6
DZF	7	1	Protein kinase	6	6
Protein kinase	7	6	Zinc-hook	4	1
EF-hand	6	3	ADF-H	3	1
GYF	5	2	EF-hand	3	1
PH (Pleckstrin homology)	5	4	GYF	3	1
ADF-H	4	1	PH	3	3
FERM	4	3			
PDZ	4	2			
Zinc-hook	4	1			

Table 8. Association of domains with CSAHs in SwissProt70 proteins

	Number of proteins			
Domain	**Observed**	**Expected**	**Observed/Expected**	**P-value**
GYF	3	0.02	171.17	7.09E-07
Zinc-hook	4	0.04	114.11	4.86E-08
Myosin_head-like	6	0.12	51.87	2.44E-09
ADF-H	3	0.06	51.35	2.96E-05
PCI	8	0.21	37.83	6.31E-11
IQ	8	0.25	31.99	2.39E-10
RRM	9	0.79	11.38	1.38E-07
PH	3	0.6	5.01	2.30E-02
EF-hand	3	0.76	3.94	4.20E-02
Protein_kinase	6	2.48	2.42	3.98E-02

Only domains associated with CSAHs in at least 3 proteins are listed.

Table 9. Annotated isoforms of CSAH-containing proteins in the full human and mouse proteomes**.**

Proteins with isoforms not affecting the CSAH region are not shown.

ID	Name	Number of isoforms	CSAH-containing isoform(s)	Length of CSAH(s)	Remark
HUMAN proteome					
TPM1_HUMAN	tropomyosin alpha-1 chain	6	6	44	no CSAH in isoforms 1-5
AFAD_HUMAN	afadin	5	2,4	82	no CSAH in isoform 1
			3,5	71	shorter version of CSAH present in these isoforms
CALD1_HUMAN	caldesmon	5	1	134, 67	2 CSAHs in isoform 1
			2,3,4,5	67	only one CSAH present in isoform 2
AKD1_HUMAN	Adenylate kinase domain-containing protein 1	6	3,4,6	41	no CSAH in isoforms 1,2,5
PSPC1_HUMAN	Paraspeckle component 1	2	2	41	no CSAH in isoform 1
MA7D2_HUMAN	MAP7 domain-containing protein 2	3	1,2	127	no CSAH in isoform 3
REN3A_HUMAN	Regulator of nonsense transcripts 3A	3	1,2	32	no CSAH in isoform 3
RENT2_HUMAN	Regulator of nonsense transcripts 2	2	1	68	no CSAH in isoform 2
ARGL1_HUMAN	Arginine and glutamate-rich protein 1	2	1	87,81	2 CSAHs in isoform 1
			2	81	only one CSAH present in isoform 2
PCLO_HUMAN	Protein piccolo	6	1,2,4,5,6	40	no CSAH in isoform 3
MOUSE proteome					
CSPP1_MOUSE	Centrosome and spindle pole associated protein 1	6	1,2,4	34	no CSAH in isoforms 3,5,6
ARGL1_MOUSE	Arginine and glutamate-rich protein 1	2	1	87,81	2 CSAHs in isoform 1
			2	81	only one CSAH present in isoform 2
KDM2B_MOUSE	Lysine-specific demethylase 2B	4	1,2	43	no CSAH in isoforms 3,4
MA7D2_MOUSE	MAP7 domain-containing protein 2	3	1,2	71	no CSAH in isoform 3
PERQ2_MOUSE	PERQ amino acid-rich with GYF domain-containing protein 2	2	1	168	no CSAH in isoform 2
SH24B_MOUSE	SH2 domain-containing protein 4B	2	1	61	no CSAH in isoform 2
NONO_MOUSE	Non-POU domain-containing octamer-binding protein	2	1	57	no CSAH in isoform 2

Data are compiled using the human and mouse proteomes with variants as supplied with UniProt 2011_05. There were 18 proteins in the human and 13 in the mouse proteomes with annotated variants that do not affect the existence or length of CSAHs.

Table 10. Annotated variants and residue modifications affecting CSAH segments

Data refer to all CSAH-containing proteins identified in full SwissProt.

ID	CSAH	site of variation / modification	description	CSAH in variant
variants				
IF2P_HUMAN	342-382	360	R -> G (in dbSNP:rs3205296).	342-392
MAP7_HUMAN	486-564	526	R -> P (in dbSNP:rs35107962).	no CSAH
RB12B_HUMAN	548-880	605	R -> C (in dbSNP:rs17857188).	548-880
TNNT2_HUMAN	154-185	170	Missing (in CMH2).	no CSAH
TNNT2_HUMAN	154-185	173	E -> K (in CMH2).	no CSAH
TNNT2_HUMAN	154-185	170 + 173	both variations in CMH2	no CSAH
TOLA_HAEIN[1]	166-233	190	A -> R (in strain: 1479).	166-233
TOLA_HAEIN	166-233	203	V -> A (in strain: 1479).	166-233
TOLA_HAEIN	166-233	227	D -> A (in strain: 1479).	166-229
TOLA_HAEIN	166-233	232	A -> AKAAAEAKAKA (in strain: 1479).	166-251
TOLA_HAEIN	166-233	all four above	strain 1479	166-251
TRHY_HUMAN[1]	493-563	552	R -> S (in dbSNP:rs6680692).	CSAH missing
TRHY_HUMAN	1395-1458	1400	R -> P (found in a renal cell carcinoma)	CSAH missing
residue modifications[2]				
PEPL_MOUSE	1299-1340	1329	Phosphoserine (By similarity)	1299-1340
RB12B_HUMAN	548-880	575	Phosphoserine	548-880[3]
RB12B_HUMAN	548-880	591	Phosphoserine	548-880[3]
RB12B_PONAB	548-621	575	Phosphoserine (By similarity)	548-621
SFPQ_HUMAN	538-599	571	Dimethylated arginine	no change expected[4]
SFPQ_MOUSE	530-591	563	Dimethylated arginine (By similarity)	no change expected[4]
SRP2_SCHPO	225-296	276	Phosphoserine	225-296[5]
SRP2_SCHPO	225-296	294	Phosphoserine	225-299[6]
SRP2_SCHPO	225-296	296	Phosphoserine	225-295[7]

[1]TOLA_HAEIN and TRHY_HUMAN contain multiple CSAH segments, only those affected by the variations are shown.
[2]In order to estimate the effect of phosphorylation, serines were replaced by glutamates in the sequences
[3]The same CSAH is predicted when both serines are phosphorylated
[4]Dimethylated arginine has the same charge (+1) as arginine [18], thus neither algorithms are expected to predict different CSAHs.
[5]The same CSAH is predicted when both Ser294 and Ser296 or all three serines are phosphorylated.
[6]The same CSAH is predicted when both Ser276 and Ser294 are phosphorylated
[7]The same CSAH is predicted when both Ser276 and Ser296 are phosphorylated

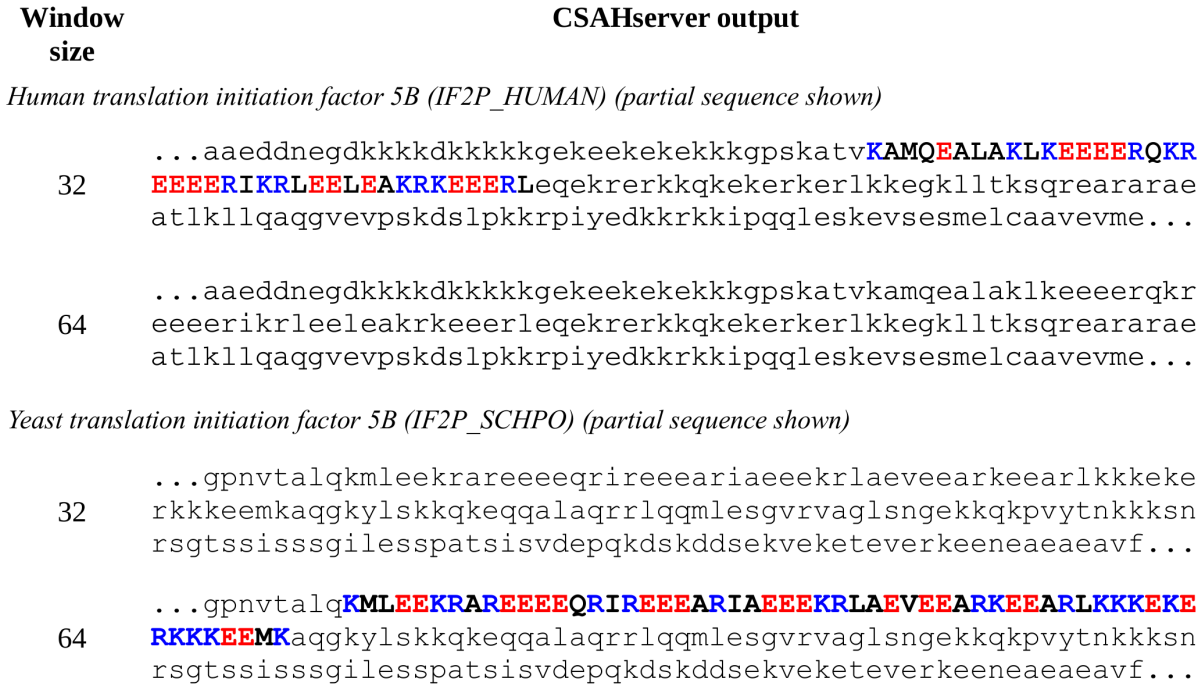
Window size
CSAHserver output
Human translation initiation factor 5B (IF2P_HUMAN) (partial sequence shown)
...aaeddnegdkkkkdkkkkkgekeekekekkkgpskatvKAMQEALAKLKEEEERQKR
32
EEEERIKRLEELEAKRKEEERLeqekrerkkqkekerkerlkkegkllтksqreararae
atlkllqaqgvevpskdslpkkrpiyedkkrkkipqqleskevsesmelcaavevme...
...aaeddnegdkkkkdkkkkkgekeekekekkkgpskatvkamqealaklkeeeerqkr
64
eeeerikrleeleakrkeeerleqekrerkkqkekerkerlkkegkllтksqreararae
atlkllqaqgvevpskdslpkkrpiyedkkrkkipqqleskevsesmelcaavevme...
Yeast translation initiation factor 5B (IF2P_SCHPO) (partial sequence shown)
...gpnvtalqkmleekrareeeeqrireeeariaeeekrlaeveearkeearlkkkeke
32
rkkkeemkaqgkylskkqkeqqalaqrrlqqmlesgvrvaglsngekkqkpvytnkkksn
rsgtssisssgilesspatsisvdepqkdskddsekveketeverkeeneaeaeavf...
...gpnvtalqKMLEEKRARЕЕЕЕQRIREEEARIAEEEKRLAEVEEARKEEARLKKKEKE
64
RKKKEEMKaqgkylskkqkeqqalaqrrlqqmlesgvrvaglsngekkqkpvytnkkksn
rsgtssisssgilesspatsisvdepqkdskddsekveketeverkeeneaeaeavf...

Figure 1.

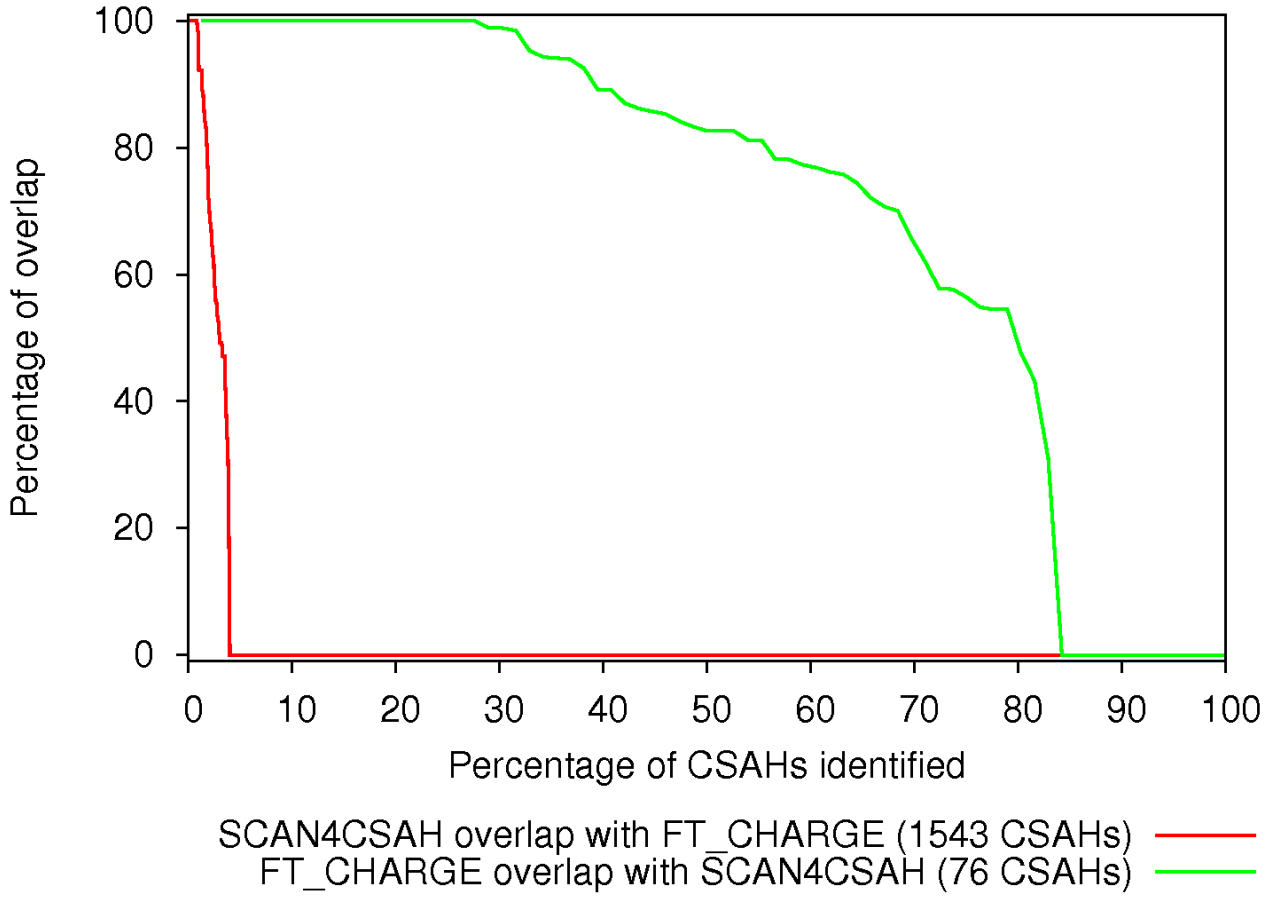
Percentage of overlap
100
80
60
40
20
0
0
10
20
30
40
50
60
70
80
90
100
Percentage of CSAHs identified
SCAN4CSAH overlap with FT_CHARGE (1543 CSAHs)
FT_CHARGE overlap with SCAN4CSAH (76 CSAHs)

Figure 2

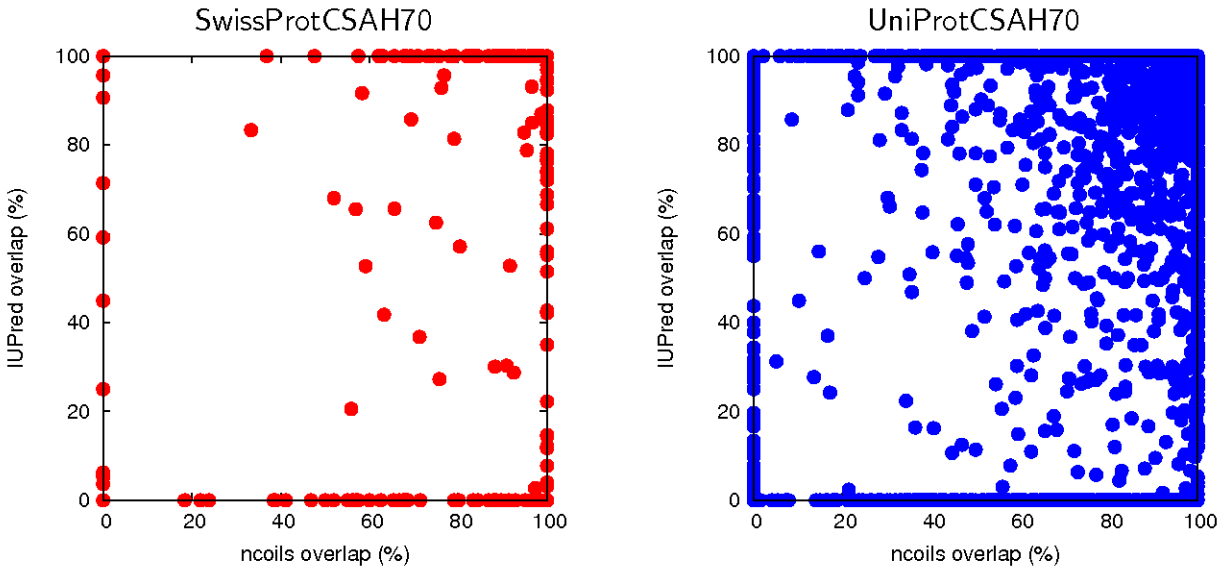
SwissProtCSAH70
IUPred overlap (%)
ncoils overlap (%)
0
20
40
60
80
100
UniProtCSAH70
IUPred overlap (%)
ncoils overlap (%)

Figure 3

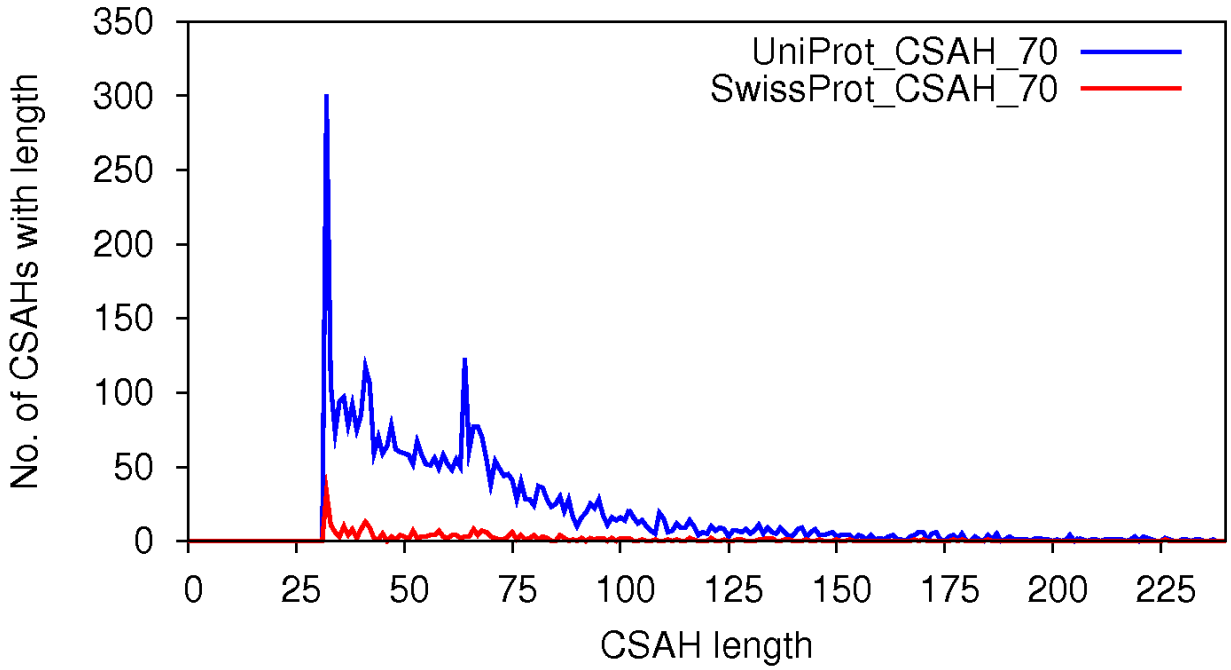
UniProt_CSAH_70
SwissProt_CSAH_70
No. of CSAHs with length
CSAH length
0
50
100
150
200
250
300
350
0
25
50
75
100
125
150
175
200
225

Figure 4

Figure 5


MYO6_HUMAN
MYO6_PIG
MYO6_MOUSE
MYO6_CHICK
MYS9_DROME
MYO1_COPC7
MYO1_NEUCR
MYO1_MAGO7
MYO1_SCHPO
MYO10_BOVIN
MYO10_HUMAN
MYOI_DICDI
MYS2_DICDI
MYO1_CAEEL
MYO1_YEAST
MYOM_DICDI

Figure 6

```
MYO1O_BOVIN RRRFLHLKKAAVVFQKQLRGQIARRVYRQLLAEKRAEEEKRKREEEEKRKREEEERERERERREAELRAQQEEAARKQRELEALQQESQRAAELSRELEKQ
MYO1O_HUMAN RRRFLHLKKAAIVFQKQLRGQIARRVYRQLLAEKREQEEKKKQEEEEKKKREEEERERERERREAELRAQQEEETRKQQELEALQK-SQKEAELTRELEKQ
```

Figure 7

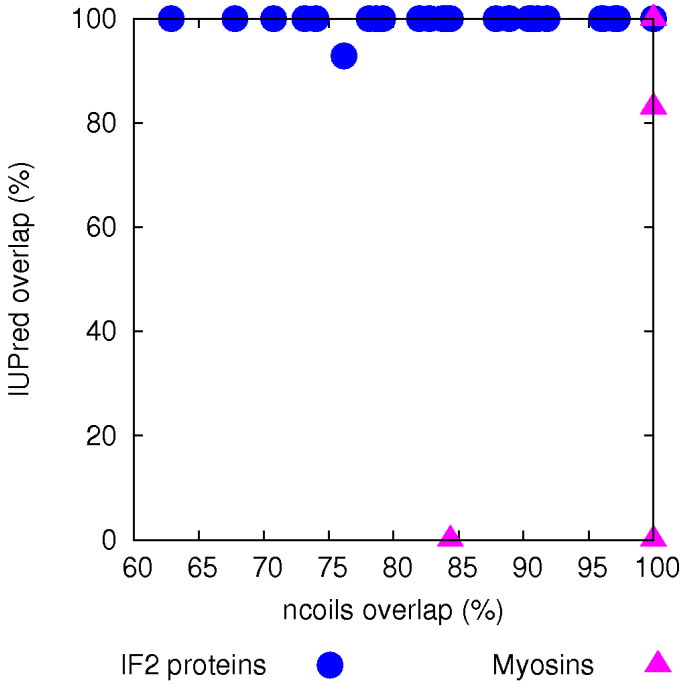
100
80
60
40
20
0
IUPred overlap (%)
60
65
70
75
80
85
90
95
100
ncoils overlap (%)
IF2 proteins
Myosins

Figure 8