PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Charged single α-helix: A versatile protein structural motif

Dániel Süveges,[1†] Zoltán Gáspári,[2†] Gábor Tóth,[3] and László Nyitray[1*]

[1] Department of Biochemistry, Eötvös Loránd University, Pázmány Péter s. 1/C, H1117 Budapest, Hungary
[2] Institute of Chemistry, Eötvös Loránd University, Pázmány Péter s. 1/C, H1117 Budapest, Hungary
[3] Bioinformatics Group, Agricultural Biotechnology Center, Szent-Györgyi Albert u. 4, H2100 Gödöllő, Hungary

ABSTRACT

A few highly charged natural peptide sequences were recently suggested to form stable α-helical structures in water. In this article we show that these sequences represent a novel structural motif called "charged single α-helix" (CSAH). To obtain reliable candidate CSAH motifs, we developed two conceptually different computational methods capable of scanning large databases: SCAN4CSAH is based on sequence features characteristic for salt bridge stabilized single α-helices, whereas FT_CHARGE applies Fourier transformation to charges along sequences. Using the consensus of the two approaches, a remarkable number of proteins were found to contain putative CSAH domains. Recombinant fragments (50–60 residues) corresponding to selected hits obtained by both methods (myosin 6, Golgi resident protein GCP60, and M4K4 protein kinase) were produced and shown by circular dichroism spectroscopy to adopt largely α-helical structure in water. CSAH segments differ substantially both from coiled-coil and intrinsically disordered proteins, despite the fact that current prediction methods recognize them as either or both. Analysis of the proteins containing CSAH motif revealed possible functional roles of the corresponding segments. The suggested main functional features include the formation of relatively rigid spacer/connector segments between functional domains as in caldesmon, extension of the lever arm in myosin motors and mediation of transient interactions by promoting dimerization in a range of proteins.

Proteins 2009; 74:905–916.

Key words: structure prediction; circular dichroism spectroscopy; salt bridges; Fourier transformation; intrinsically disordered proteins; coiled coil; myosin; protein kinase.

INTRODUCTION

The vast majority of known protein structures contain primarily one or both of the two most common secondary structural elements, α-helix and β-sheet. Investigating the relationship between amino acid sequence and local structure led to the recognition of secondary structure-forming preferences of individual amino acid types.[1,2] Although a number of practically fully α-helical proteins and small peptides have been described, single α-helices stable in water in the absence of helix-promoting (co)solvents (such as 2,2,2-trifluoroethanol) have not been found in nature until very recently.[3–6] Helices in proteins form extensive interactions with other parts of the chain (so-called supersecondary elements: i.e. α-β hairpins, helical bundles, or coiled coils), or in membrane proteins they interact fully or partially with membrane lipids. Model peptides adopting single α-helical conformation were successfully designed.[7–10] These peptides contained residues with high helix-forming potential, for example, alanine and several charged residues necessary for stabilization in water. It was speculated that introducing *i*, *i*+4 or *i*, *i*+3 salt bridges adds extra stabilization to these helices.[11] Another proposed feature of these sequences was that charged side-chains stabilize helices by sequestering water molecules away from the protein backbone, thereby promoting intramolecular hydrogen bond formation.[12,13] Hydrophobic interaction between nonpolar side-chains and/or alkyl groups of long chain charged groups could also contribute to helix stability.[14]

The handful of segments from natural proteins known to form stable single α-helical structure in water, such as in caldesmon,[5,6] consist mainly of charged residues, and are thought to be stabilized by *i*, *i*+4 salt bridges. Interestingly, three myosin families (myosin 6, myosin 10, and myosin 7) contain highly charged sequences in their tail regions that were originally predicted to form coiled coils, but later

Additional Supporting Information may be found in the online version of this article.
Abbreviations: CD, circular dichroism; CSAH, charged single α-helix; IDP, intrinsically disordered proteins.

Grant sponsor: Hungarian Scientific Research Fund (OTKA); Grant numbers: 61784, 49812, 68466, and 68079.
[†]Dániel Süveges and Zoltán Gáspári contributed equally to this work.
*Correspondence to: László Nyitray, Department of Biochemistry Eötvös Loránd University, Pázmány Péter s. 1/C, H1117 Budapest, Hungary. E-mail: nyitray@elte.hu
Received 31 March 2008; Revised 21 May 2008; Accepted 9 June 2008
Published online 19 August 2008 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.22183

shown to form, in the case of myosin 10, stable single α-helix with a characteristic charge distribution.[3] This segment in myosin 10 was suggested to lengthen the lever arm of the myosin motor and hence to increase the size of the power stroke.[3]

To address the question whether these sequences represent a previously unrecognized protein structural motif, termed here charged single α-helix (CSAH), we designed two independent algorithms, SCAN4CSAH and FT_CHARGE, capable of detecting candidate segments in protein sequence databases. On the basis of the consensus hits of the two approaches, we selected, cloned and expressed three protein segments. Using circular dichroism (CD) spectroscopy, we found that these segments indeed form stable α-helices in water; thus, the existence of the charged single helix motif, as a widely occurring linear structural element could be confirmed.

MATERIALS AND METHODS

The SCAN4CSAH method

SCAN4CSAH is based on sequence features described earlier as being characteristic for salt bridge-stabilized single α-helices. The method allows the weighted quantitative scoring of various types of interactions that may affect the stability of a putative CSAH motif either positively or negatively. The stabilizing interactions and modifying effects considered are as follows: (1) salt bridges between oppositely charged amino acid residues three or four positions apart (i, i+3 or i, i+4; e.g. KxxxE)[3,6,11,15,16]; (2) the cooperative stabilizing effect of three consecutive, alternately charged amino acids at appropriate positions (e.g. i, i+4, i+8 or i, i+4, i+7 or i, i+3, i+7; like in ExxxRxxxE)[11,15]; (3) the relative orientation of the oppositely charged residues (the ionic pair may be more stabilizing if the acidic residue is in N-terminal position).[15] The helix-destabilizing effects of both attracting and repulsing interactions are taken into account for oppositely charged side chains spaced i, i+1 or i, i+2 as well as for residues with identical charges spaced i, i+4 or i, i+3 in the sequence.[16]

We implemented the method in a Perl program that counts the numbers of putative i, i+4 and i, i+3 stabilizing interactions and the 3-residue combinations thereof. Ionic pairs are counted separately for the two possible orientations (i.e. acidic or basic N-terminal residue). The four possible destabilizing pairs are also counted. Each putative CSAH motif is delimited as follows: the region must contain a series of consecutive stabilizing pairs with a maximum gap of five residues allowed, where a gap consists of residues that have no attracting ionic pair. The minimum length of a putative CSAH is 40 amino acids. A score is assigned to each putative CSAH: a weighted sum of all counted interactions is divided by the length of the CSAH segment to obtain the score. This relative score can be considered as a rough quantitative measure of helix stability and is used to rank putative CSAH motifs. We use such length-normalized score to avoid false positives that would otherwise arise from long albeit weakly charged regions. Thus, the 40-amino acid minimum length threshold is needed to exclude less significant short CSAH-like segments.

The weighting factors or weights assigned to the different types of interactions are positive or negative numbers depending on whether the particular interaction is thought to exert a positive or negative effect on helix stability, respectively. Estimation of the weights is described in detail in Supplementary Methods.

We scanned the Swiss-Prot (Release 54.2) protein sequence database for putative CSAH-containing entries. Most proteins did not contain any putative CSAH that met the 40-amino acid minimum length requirement and thus received no CSAH score at all. Proteins with one or more putative CSAH motifs received the score of the highest scoring putative CSAH found in that protein. 8250 of the 283454 Swiss-Prot proteins could be scored. Significance of a CSAH motif or a CSAH-containing protein was estimated by comparing its score to the distribution of scores obtained for the scorable database entries. Score distribution of the scorable proteins can therefore be approximated by an extreme value distribution (see Supplementary Methods for details). A generalized extreme value (GEV) distribution was fit by maximum likelihood using the R package "evd".[17,18] The significance level of an observed score can be calculated from the fitted analytical distribution. The P-value high threshold is set to $P = 0.01$, resulting in 143 Swiss-Prot proteins accepted as CSAH-containing ones by the SCAN4CSAH method.

The FT_CHARGE method

The method presented below bears similarities to other FT-based applications on biopolymer sequences, for example spectral repeat finder.[19] Charges can be assigned as follows: −1 for Asp and Glu, +0.5 for His, +1 for Arg, and Lys and zero for any other amino acid residue. Fourier transformation can be applied directly to these charges or the charge correlation function R defined for the sequential distance n as:

$$R(n) = \sum_{i=1}^{m-n} c(i)c(i+n) \qquad (1)$$

where $c(i)$ is the charge assigned to the ith residue of the sequence, and m is the length of the sequence.

This function is sensitive to the characteristic distances of identically and oppositely charged residues and gives higher inherent sensitivity for charged patterns than using the charges in the sequence alone. The output of FT is a spectrum with amplitudes as a function of fre-

quencies. Repeatedly charged sequences such as CSAHs are expected to yield high amplitudes at certain frequencies corresponding to amino acid distances along the sequence. As amplitudes are arbitrary, and, more importantly, are dependent on the inherent charge density of the sequence, no absolute threshold can be established as indicative of CSAHs. Therefore, a statistical approach based on the randomization of the input sequence can be used. In this case, the charge correlation function and its FT are computed for a chosen number of random sequences composed of the same residues as the original. Thus the periodicity, if present, is effectively eliminated from the random sequences while maintaining the same overall charge density. The relevance of the observed amplitudes in the native sequence can be assessed using a *Z*-score.

$$Z = \frac{A_{\max}^{\text{native}} - \overline{A_{\max}^{\text{random}}}}{\sigma_{\text{random}}} \quad (2)$$

where $A_{\max}$ denotes the maximum amplitude found in the FFT spectrum and σ the standard deviation of the maximum amplitudes in the random sequences. The bar denotes average over all randomly generated sequences.

CSAHs can be identified as sequences having high *Z*-scores and a periodicity corresponding to 6–9 amino acid residues (corresponding to the maximum amplitude). In practice, sequence segments are scanned instead of full sequences. For FFT (Fast Fourier Transform) to be applied, segments with lengths corresponding to different powers of 2 (e.g. 64) are used and segments are generated in an overlapping manner (e.g. a 64-residue segment overlaps 32 residues with the previous and the next one). For database searches, nine random sequences plus the original were used to obtain the average maximum randomized amplitude for *Z*-score calculation. The method FT_CHARGE was implemented in Perl.

Database searches

The Swiss-Prot database[20] (Release 54.2) was downloaded from the Expasy website (http://www.expasy.org) and searched for candidate CSAH motifs by both SCAN4CSAH and FT_CHARGE (using 64-residue segments with 32-residue overlaps). SCAN4CSAH results were ranked by the maximum relative score for each protein and FT_CHARGE hits were ranked by a combined score obtained by multiplying the amplitude with the *Z*-score after frequency filtering (hits in the frequency range of 1/10 to 1/6 were retained).

Expression and purification of potential α-helical protein segments

Three candidates were chosen to validate the predictions: MYO6 (Q9UM54, residues: 933–985), GPC60 (Q9H3P7, residues: 183–238), and M4K4 (O95819, residues: 417–480). PCR products were ligated into pET-15b (Novagen) vector between NdeI and BamHI sites. After verifying the sequences, constructs were transformed into BL21 (DE3) pLysS (Novagen) cells. After induction with 1 m*M* IPTG (Sigma), cells were grown at 37°C for 3 h in 2YT medium. His-tagged proteins were purified on Ni^{2+}-affinity column (Novagen). After removal of the His-tags by thrombin (Sigma) the recombinant peptides were further purified by ion exchange chromatography, using Amersham HiTrap Q HP and SP HP columns (depending on the isoelectric points of the peptides).

Circular dichroism spectroscopy

CD spectra were recorded on a Jasco J720 spectropolarimeter, using 1-mm cuvette with wavelength range between 190 and 260 nm; in 20 m*M* phosphate buffer pH 7.5, 50 m*M* NaCl with protein concentration between 10 and 50 μ*M*. The temperature-controlled cuvette was maintained at 20°C with a Neslab RTE-111 circulating water bath. Results are expressed in mean residual molar ellipticity $[\Theta]$ (deg cm^2 $dmol^{-1}$) calculated from Eq. (3).

$$[\Theta] = \frac{\Theta_{\text{obs}}\text{MRW}}{10lc} \quad (3)$$

where Θ_{obs} is the observed ellipticity expressed in millidegrees, MRW is the mean residue molar weight (molecular weight of the proteins divided by the residue number), l is the pathlength in cm, and c is the protein concentration expressed in mg/mL.

Spectra were analyzed using the CDSSTR[21] program available on DICHROWEB[22] with reference set 4 to calculate the fraction of α-helix. To determine the effect of temperature on helical content, the ellipticity at 222 nm was monitored upon continuous heating with a rate 1°C/min in the temperature range 5–80°C. Dependence of the structural stability of MYO6 and M4K4 on ionic strength was measured by CD spectroscopy recording the ellipticity at 222 nm as a function of NaCl concentration. The peptides were analyzed in 10 m*M* sodium phosphate buffer (pH 7.0) at 20°C at a peptide concentration of 10 μ*M* and at various NaCl concentrations. The concentration dependence of CSAH peptides were measured in 20 m*M* sodium phosphate buffer (pH 7.0) with 50 m*M* NaCl at 20°C, at a range of protein concentration from 10 to 600 μ*M*. Spectra were recorded from 260 to 190 nm. The pathlength of the cuvettes were chosen adequately to each protein concentration applied.

Chemical crosslinking of CSAH peptides

To assess the multimeric nature of CSAH peptides chemical crosslinking experiments were carried out using dimethyl pimelimidate. The reaction buffer contained 50 m*M* sodium phosphate (pH 8.0) with 50 m*M* sodium chloride, the protein concentrations were approximately 150 μ*M*, the reaction was started by adding 3 m*M* cross-

Table I
Selected CSAH Motifs Identified by Both SCAN4CSAH and FT_CHARGE within the First 100 Hits

SCAN4CSAH rank number	FT_CHARGE rank number	Swiss-Prot Entry name[a]	Swiss-Prot ID	SEGMENT	SEQUENCE[b]
1	1	MNN4_YEAST	P36044	1031–1173	AKLLEERKRREKKKKEEEEKKKKEEEEKKKKEEEEKKKKEEEEKK KKEEEEKKKKEEEEKKKQEEEEKKKKEEEEKKKQEEGEKMKNE DEENKKNEDEEKKKNEEEEKKKQEEKNKKNEDEEKKKQEEEE KKKNEEEEKKKQE
3	47	**GCP60_HUMAN**	Q9H3P7	179–242	HKIEKEEQEKKRKEEEERRRREEEERERLQKEEEKRRREEEERLR REEEERRRIEEERLRLEQQ
8	25	**MYO6_HUMAN**	Q9UM54	915–983	KKKQQEEEAERLRRIQEEMEKERKRREEDEKRRRKEEEERRMK LEMEAKRKQEEEERKKRED DEKRIQA
19	16	PERQ2_HUMAN	Q6Y7W6	729–896	LEKAKAAKLEQERRRAEMRAKRREEEERKRQEELRRQQEEILRR QQEEERKRREEEELARRKQEEALRRQREQEIALRRQREEEERQ QQEEALRRLEERRREEEERRKQEELLRKQEEEAAKWAREEEE AQRRLEENRLRMEEEAARLRHEEEERKRKELEVQRQKELM
20	70	CALD1_RAT	Q62736	304–370	LEELKKKREERRKVLEEEEQRRKQEEADRKAREEEEKRRLKEEI ERRRAEAAEKRQKMPEDGLSEDK
28	80	NONO_PONPY	Q5RFL9	314–370	QDLMRRQEELRRMEELHNQEVQKRKQLELRQEEERRRREEEM RRQQEEMMRRQQEGF
46	53	**M4K4**_HUMAN	O95819	377–479	QEQQLREQEEYKRQLLAERQKRIEQQKEQRRRLEEQQRREREA RRQQEREQRRREQEEKRRLEELERRRKEEEERRRAEEEKRRVE REQEYIRRQLEEEQRHL
47	49	AFAD_HUMAN	P55196	1578–1659	QRLEAERRARLQDEERRRQQQLEEMRKREAEDRARQEEERRRQ EEERTKRDAEEKRRQEEGYYSRLEAERRRQHDEAARRLL
48	67	GLE1_YEAST	Q12315	150–230	RENKRRVEEQRKRKEEEERKRKEAEEKAKREQELLRQKKDEEE RKRKEAEAKLAQQKQEEERKKIEEQNEKERQLKKEHEA
72	88	IF2_RHIME	Q92SW4	114–178	GEMEARRRALAEAQIRDAEEAKRRAEDEVRRRREEEERLAREK EEAARRAAEEAARPPVEAEKTE

[a]Sequences analyzed by CD spectroscopy are boldface.
[b]Negatively and positively charged residues are underlined and boldface, respectively.

linker. After 60 min incubation at room temperature, the reaction was quenched by adding 1 volume of SDS-PAGE Loading buffer (125 m*M* Tris pH 6.5, 8% SDS, 712 m*M* 2-mercaptoethanol). 2.5 μL of each reaction was loaded on a 17% SDS-PAGE.

Computational analysis of candidate proteins with CSAH motif

To assess the role and significance of the CSAH motif in proteins, consensus hits from the first 100 proteins detected by both methods were analyzed in detail. In particular, coiled coils and intrinsically disordered regions were searched for using the programs MultiCoil,[23] Paircoil2,[24] Amphisearch[25] Marcoil,[26] COILS,[27,28] and IUPred.[29,30] Functional information and domain organization was derived from SwissProt[20] and the Pfam databases.[31]

RESULTS

Performance and comparison of the two CSAH detection methods

The SCAN4CSAH approach is based on sequence features suggested earlier as being characteristic for a stabilized single α-helix where intrachain ionic interactions are considered to be primarily responsible for helix stability. FT_CHARGE identifies regions with single α-helix forming potential by applying Fourier transformation (FT) to the charges assigned to residues along the sequence (for details see Materials and Methods). The application of two independent approaches is expected to enhance the reliability of CSAH detection. Indeed, all consensus hits investigated experimentally showed high α-helical content (see below). The fact that the hits identified by both algorithms (full list of consensus hits is shown in Supplementary Table I) contain a number of close homologs also reinforces the reliability of the consensus approach.

The two methods presented here capture different aspects of sequence motifs with CSAH-forming potential. Although the score-based algorithm was optimized to find segments known or strongly suspected to form CSAHs, the FT-based one is an ab initio method with no built-in bias to recognize selected targets. FT_CHARGE is able to identify patterns with a variety of frequency values. Thus, periodicities characteristic of possible charged helices should be selected for comparison. Considering that α-helices have 3.6 residues per turn, periodicity values around 1/7 are expected to be meaningful. This reasoning is in good agreement with empirical anal-

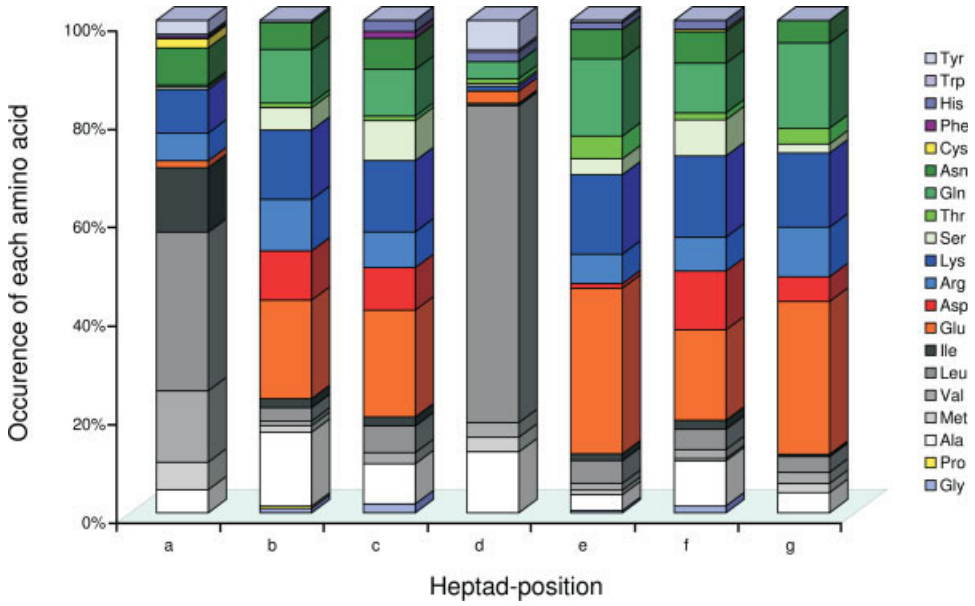

Figure 1

Amino acid composition of experimentally verified coiled coil sequences for each heptad position. The prevalence of charged amino acids except for the *a* and *d* positions is apparent. The coiled coil predictor Marcoil, and its training database, with 419 annotated coiled-coil sequences were used for the analysis.

ysis of the hits, as previously suggested CSAH segments can be observed with frequencies 1/10 to 1/6. Using these criteria the two methods gave reasonably similar results: the first hit in the Swiss-Prot[20] search with both methods is the yeast MNN4 protein, exhibiting an almost perfect pattern of four positively charged residues followed by four negatively charged ones spanning almost 60 residues. Within the first 100 hits, the two methods yield 47% overlap (Supplementary Fig. 1). Segments identified by both methods overlap in average for about 53 residues, corresponding to more than 83% of the window length used in FT_CHARGE (64 residues). In light of the difference between the two approaches, the order of the hits is expected to be only approximately the same and more dissimilar when more hits are considered, hence the above overlap can be regarded as excellent agreement between the two methods.

It is important to note that the CSAH sequences could often be detected as positive hits by coiled coil and disorder prediction programs. Of the selected proteins shown in Table I, 88% of CSAH residues were predicted to form coiled coils by the program COILS[27,28] and 67% were marked disordered by IUPred.[29,30] [We also note that coiled coil segments can often be recognized as intrinsically disordered sequences and vice versa. In DisProt release 3.6, COILS predicted 28% of the disordered residues to be in coiled coil (in 64 out of 472 proteins), whereas in 10 known coiled coil proteins taken from the PDB, 40% of the residues predicted to be coiled coils by COILS were also predicted to be disordered by IUPred.] A more detailed analysis considering all consensus CSAH sequences revealed that four of the five coiled coil prediction algorithms predicted over 70% residues to form coiled coils (involving more than 85% of the sequences, Table II). However, the fifth algorithm, Amphisearch[25], predicted no coiled coils in any of the single helices proposed. This is not surprising since this method uses thermodynamic parameters to evaluate the propensity of an amino acid sequence to form coiled coils and is biased against charged residues in the heptade core positions, *a* and *d*. The reason why the other programs mispredict CSAH regions to be coiled coils could be explained by the fact that COILS and related programs compare the query sequences to a database of known coiled coils and then derive a similarity score. Analysis of the coiled coil

Table II

Coiled Coil Prediction of CSAH Regions Using Various Predictors

	Sequences			Residues		
Predictor	Submitted	Coiled coil	Ratio	Submitted	Coiled coil	Ratio
Amphisearch	69	0	0	5472	0	0
Coils2002	69	63	91.3	5472	3930	71.8
Marcoil	69	66	95.7	5472	4133	75.5
Multicoil	69	60	87.0	5472	3905	71.4
Paircoil2	69	62	89.9	5472	4635	84.7

Table III
Results of Circular Dichroism Spectra Deconvolution

Protein	Segment	α-helix (%)	β-strand/turn (%)	Random (%)	RMSD
M4K4	417–480	70	10	13	0.008
GCP60	183–238	62	17	23	0.005
MYO6	933–985	75	13	12	0.003

regions of seven experimentally verified coiled coil containing proteins shows that, except for core positions *a* and *d*, the ratio of the charged residues is around 50% (with almost equal frequencies of positively and negatively charged residues) (see Fig. 1). This natural high charge content can bias the coiled coil predictions toward the highly charged CSAH sequences.

Selected CSAH segments form single α-helices in aqueous solution

Three CSAH candidates were chosen from among the highest scoring protein segments of the SCAN4CSAH method to check their structure experimentally. Note, that we ignored the top scoring MNN4 protein since it has regular KKKKEEEE repeats and such synthetic peptides have previously been characterized and shown to form stable α-helix in aqueous solution.[7–9] Circular dichroism (CD) spectroscopy indicates high α-helical content (62–75%, Table III) in all three recombinant pro-

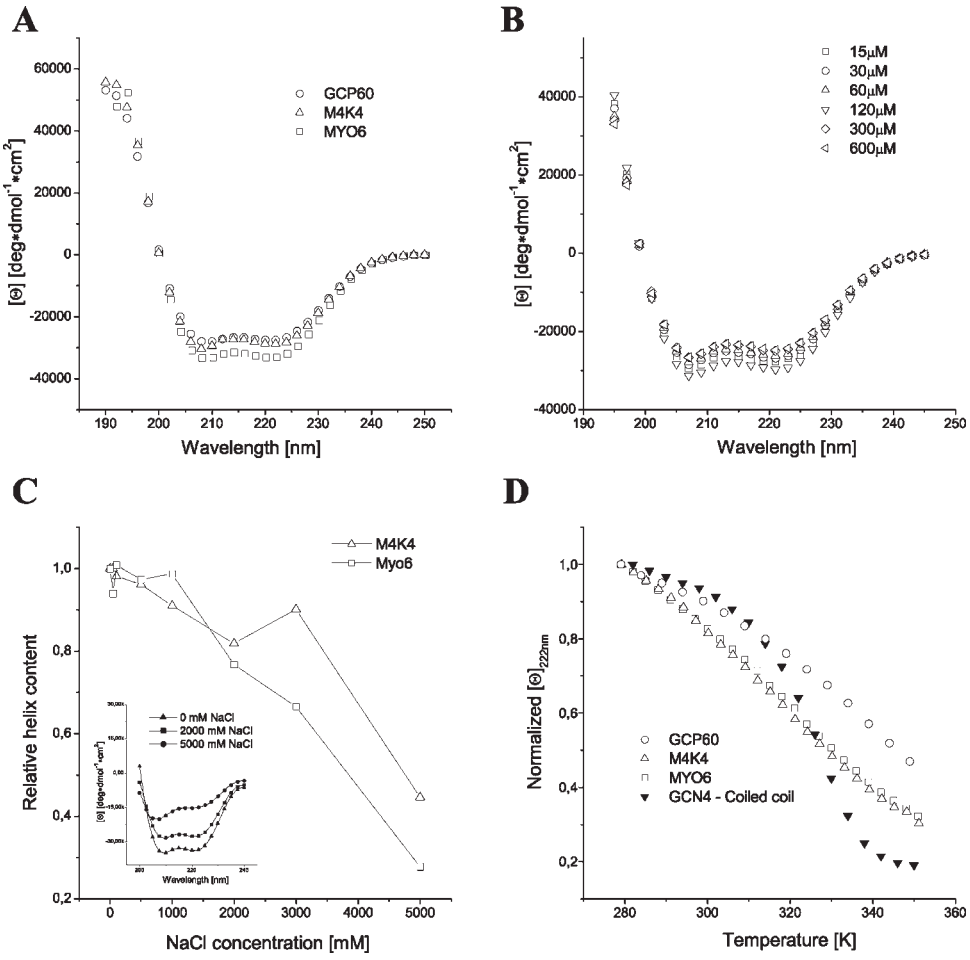

Figure 2

Circular dichroism spectroscopy of recombinant CSAH motifs. (**A**) Spectra of selected CSAH-containing peptides show high α-helical content: M4K4 70%, GCP60 62%, MYO6 75%. (**B**) Concentration dependence of the CD spectra of the CSAH region of myosin 6 (MYO6): no considerable change in the secondary structure composition can be seen in the concentration range from 15 to 600 μ*M* (all CSAH peptides examined gave similar results). (**C**) The ionic-strength dependence of the α-helix content: the CD signal (recorded at 222 nm) was normalized to the value measured without NaCl. (C Inset) Change in the CD spectra caused by increasing the salt concentration: spectra at 0 *M*, 2 *M*, and 5 *M* NaCl are shown. (**D**) Temperature denaturation: mean residual ellipticity normalized to the value measured at 6°C, as a function of temperature.[3] For comparison the coiled-coil forming leucine zipper domain of GCN4 is shown.

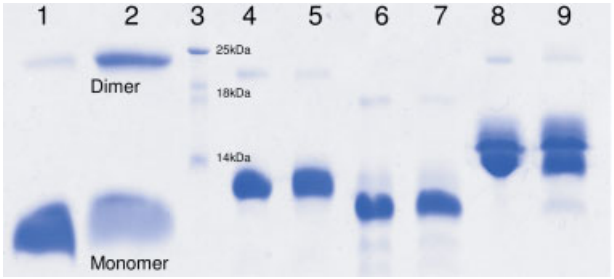

Figure 3

Chemical crosslinking of CSAH peptides followed by gel electrophoresis. The chemical crosslinking of CSAH peptides (~150 μ*M*) with dimethyl pimelimidate shows no multimerization in comparison with the coiled coil S2 domain of myosin II where the dimer formation is apparent. Lane1: S2 0 min; 2: S2 60 min; 3: marker; 4: MYO6 60 min; 5: MYO6 0 min; 6: M4K4 60 min; 7: M4K4 0 min; 8: GCP60 60 min; 9: GCP60 0 min. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

tein fragments examined [Fig. 2(A)]. Spectra recorded from 15 μ*M* to 0.6 m*M* protein concentration gave practically identical results [Fig. 2(B)] ruling out coiled coil formation or other cooperative multimerization processes. The monomeric behavior of CSAH peptides are verified by crosslinking experiments, showing no considerable increase of the dimeric fraction in comparison with the coiled coil forming subfragment 2 domain of cardiac myosin[32] (see Fig. 3). Thus, our observations unambiguously demonstrate that these protein segments form stable single α-helices in water. Thermal unfolding of the CSAH peptides is a noncooperative transition, without a characteristic melting point, unlike the cooperative two-state denaturation of the coiled coil forming leucine zipper of GCN4. Interestingly, the CSAH-containing peptides retain a relatively high α-helical content even at 80°C [Fig. 2(D)]. These results could be related to the observation that salt bridges are among the major factors increasing protein thermostability.[33] Increasing the salt concentration causes a decrease in the ellipticity at 222 nm indicating decreased α-helical content and increased disordered structure [Fig. 2(C)]. The above results are similar to that described for the temperature and ionic strength dependence of CSAH domains of caldesmon and myosin 10.[3,6]

Proteins with predicted CSAH motifs have a great variety of functions

Representatives of the 47 proteins predicted by both methods in the first 100 hits are listed in Table I, and domain organization of selected ones is shown in Figure 4. Besides the three proteins investigated experimentally, we have selected sequences that (*a*) are represented by more than one ortholog among the consensus hits, (*b*) had been investigated experimentally by other groups, or (*c*) were ranked relatively high by both methods. The level of structural and functional description of these proteins varies greatly. Their detailed investigation and comparison with their orthologs is expected to shed light on the possible function(s) of the CSAH motifs.

The presence of a CSAH domain in the tail region of myosin 6 (MYO6) is supported by our experimental data confirming experimental results obtained for myosin 10[3]. In dimeric myosins the heavy chains are known to dimerize via the coiled coil C-terminal tail domains. However, in the myosin classes containing a CSAH motif, we hypothesize that the CSAH segments form a transient, non coiled coil dimer [Fig. 5(A)]. When truncated mutants of porcine myosin 6 were analyzed, Park *et al.*[34] found that forced dimers remained together only if the construct contained the CSAH domain. Alternatively, the CSAH segments in the monomeric form could function as an extension of the lever arm, lengthening the step size of the motor as suggested by Knight *et al.*[3] for myosin 10 [Fig. 5(B)]. It remains to be determined whether

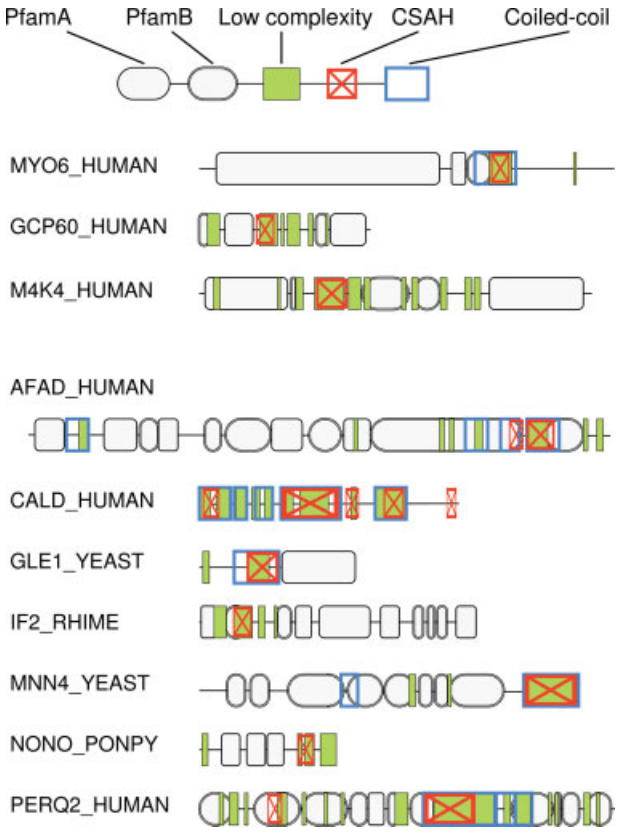

Figure 4

Domain organization of several proteins with CSAH motifs. Domain assignment is based on the Pfam annotation plus the SCAN4CSAH method. Several regions have multiple assignments; CSAHs are typically indicated as low complexity regions and, in some instances, as coiled coils. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

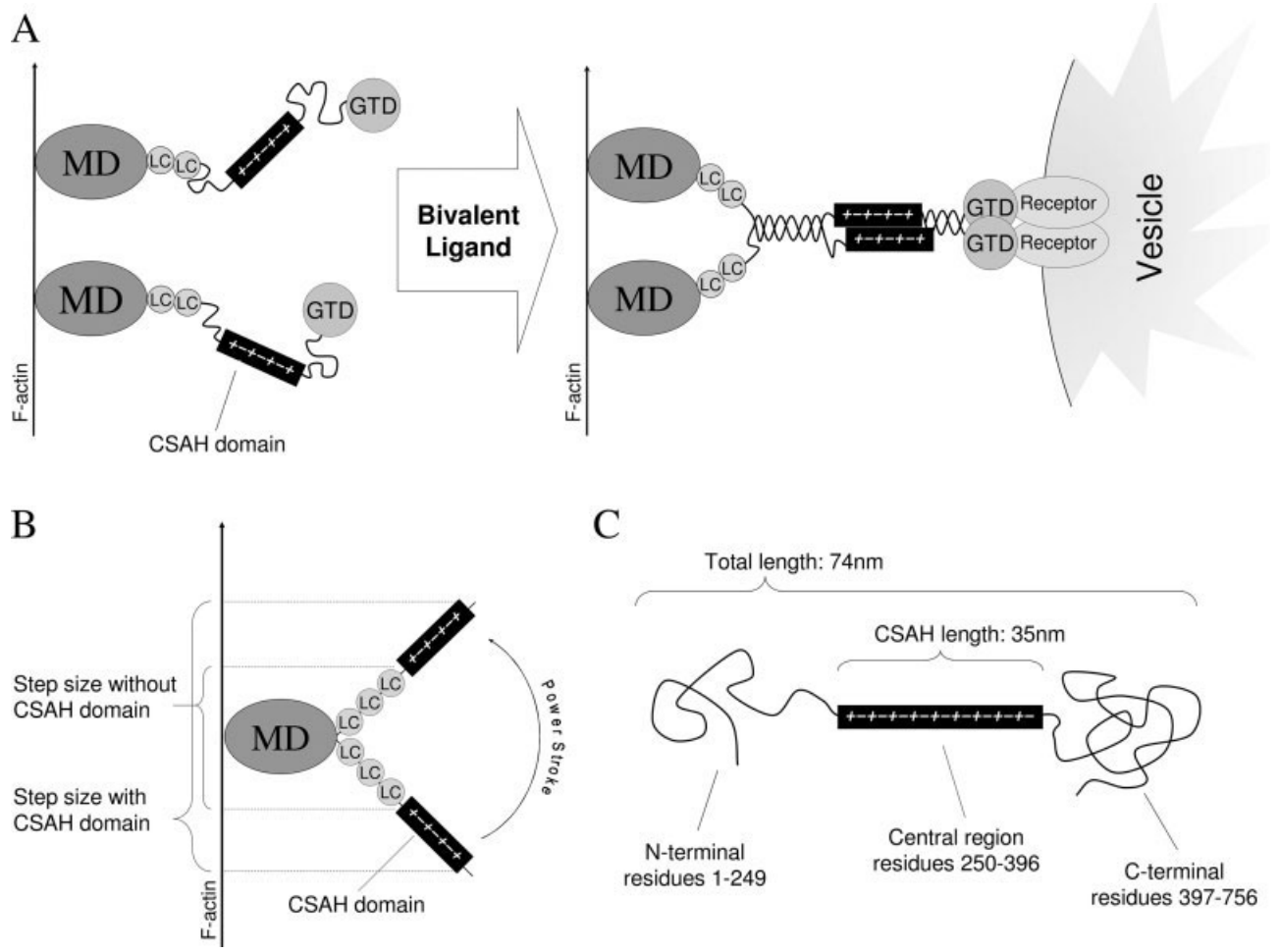

Figure 5

Schematic views of the possible roles of CSAH domains (**A**) The CSAH domain of myosin 6 could be responsible for dimerization upon binding to bivalent ligands (MD, motor domain; GTD, globular tail domain; LC, light chain).[34] (**B**) The lever arm of myosin 10 is lengthened by the CSAH domain that directly follows the light chain binding IQ motifs[3] (MD, motor domain; LC, light chain). (**C**) The CSAH domain in the central region of smooth muscle caldesmon may form a spacer between the N-terminal actin binding site and the C-terminal myosin binding region.[6,35]

these two proposed functions of the CSAH motifs in myosin motors are reconcilable. The CSAH region in myosin 6 is quite conservative from vertebrates to a Choanoflagellate (A9URT4), and most of them can be predicted by both of our programs (Supplementary Table II).

The CSAH motif of caldesmon was not only expressed and spectroscopically characterized,[6] but molecular dynamics simulations also reinforced its existence.[15] Note that the CSAH motif is only present in the smooth muscle isoform of caldesmon.[36] It is assumed that the CSAH domain acts as a spacer keeping proper distance between the N-terminal myosin and calmodulin binding domains and the C-terminal actin and tropomyosin binding regions as suggested based on EM images[37] [Fig. 5(C)].

GCP60 (Golgi complex associated protein of 60 kDa)[38] is a protein suggested to be involved in maintaining the integrity of the Golgi apparatus and in protein transport between the Golgi and the endoplasmic reticulum. Besides description of the central segment as a predicted coiled coil,[39] no hint for the role of the CSAH motif was noted.

M4K4 (mitogen-activated protein kinase kinase kinase kinase 4, also known as HGK for HPC/GCK-like kinase) plays a role in signal transduction.[40] The charged region is located C-terminal to the kinase domain and N-terminal to a segment of varying length in different isoforms. All regions C-terminal to the kinase domain are presumed to be involved in interactions with regulatory factors. Although CSAH segment could only be found in a few orthologs of M4K4, interestingly, many of them were predicted to contain a conventional coiled coil (Supplementary Table II).

The nucleoporin GLE1 is located on the cytoplasmatic side of the nuclear membrane, as a component of the nuclear pore complex, and is involved in the nucleocytoplasmatic transport. It interacts transiently with many proteins. Though the function and protein–protein interactions in yeast and human GLE1 are highly conserved,[41] regular CSAHs could only be found in the central region of fungal GLE1 whereas the orthologous domains in higher eukaryotes have a conventional coiled coil sequence (Supplementary Table II).

MNN4 is a type II membrane protein involved in mannosylphosphate transfer in yeast. The CSAH region is located intracellularly as predicted by HMMTOP.[42] This region, containing KKKKEEEE (K_4E_4) repeats, is essential since its truncation results in the *mnn4* phenotype. The closest identified MNN4 relative (Yjr061p) lacks this repeating region.[43]

Bacterial IF2 proteins are essential translation initiation factors for ribosomal protein synthesis. Functional domains of IF2 molcules are conserved throughout the three kingdoms of life.[44] Alignment of several of the hits to the core alignment given by Sorensen *et al.*[44] shows that the N-terminal highly charged segments are found in IF2s from multiple organisms including human and *Drosophila*, but the periodicities of these are not in the range characteristic of CSAH motifs (data not shown). It was noted earlier that human and yeast IF2s contain a highly charged N-terminus with low sequence similarity.[45]

Afadin is a homodimeric multidomain protein involved in cell-cell adhesion. Long (l) and short (s) splice variants of afadin exist. While s-afadin has two Ras-associated domains, a forkhead-associated domain, a DIL domain, a PDZ domain, and two proline rich domains, l-afadin continues in a CSAH domain as well as it has a third proline rich and an F-actin binding domain. L-afadin has no actin filament cross-linking activity[46] which implies that the CSAH domain may prevent it by transiently dimerizing the C-terminal regions together.

NONO (non-POU domain-containing octamer-binding protein) is a nuclearly localized protein that may form heterotetramers with splicing factor proline/glutamine-rich protein (SFPQ) and bind single-stranded RNA during mRNA maturation due to its RNA recognition motifs.[47] Some of the NONO orthologs contain a predicted conventional coiled coil region at the position of the CSAH motif in other orthologs (Supplementary Table II).

PERQ2 (also known as GIGYF2) was identified as a binding partner of Grb10, therefore its role was proposed to be in tyrosine kinase receptor signaling. Comparing the protein with its close homolog GIGYF1,[48] it is apparent that the CSAH motif is unique to PERQ2.

DISCUSSION

It has recently been recognized that stable single α-helical segments in natural sequences exist in a few proteins, including caldesmon and myosin 10. It was noted by Knight *et al.*[3] that these sequences are rare in nature, however, to our knowledge, no systematic search for such motifs has ever been performed. In this article our computational and experimental results revealed CSAH sequences in a considerable number of proteins suggesting that it is a more widely occurring structural motif. It should be noted that only the annotated Swiss-Prot database was searched by our programs, suggesting that the potential number of CSAH motif containing proteins is expected to be considerably higher than that reported in this paper. *In vitro* experiments showed that these segments form stable, monomeric α-helical structures in water at physiological conditions. The fact that similar motifs were found in a number of unrelated proteins suggests that the CSAH motif is indeed a universal structural domain, having characteristic functional role(s). Our studies also underline the necessity of using the consensus of different methods to detect sequential features of proteins.

We note that the consensus hits were almost exclusively predicted to have high (>60%) helical content by the AGADIR server, an implementation of the helix/coil transition theory.[49] Only 6 of the 69 sequences have a helical propensity below 50% and only two are estimated to have practically zero helical content. These results strengthen our conclusions regarding the stability of the CSAH sequences as single α-helices.

Our results yield clear evidence that the detected CSAH motifs are neither coiled coils nor do they correspond to intrinsically disordered protein segments. The former possibility is excluded by the temperature and concentration dependence of the CD curves and also by chemical cross-linking studies, whereas the latter can be rejected simply by the shape of the CD curves as disordered segments typically give rise to a CD signal characteristic of polyproline II conformation, believed to represent a conformational average.[50] The amino acid frequencies of CSAH motifs (as detected by SCAN4CSAH) among the common hits also diverge from that reported for intrinsically disordered proteins (IDPs),[51] the former having an excess of Glu and Arg and a lack of Pro (Fig. 6 and Supplementary Table III). These frequencies are consistent with the high charge density and helical conformation of CSAHs and represent an even more pronounced difference from globular proteins than from IDPs. The fact that CSAH motifs can be detected by specific methods together with the finding that a large portion of the corresponding segments are recognized both as coiled coils and intrinsically disordered regions by different predictors calls for critical assessments of the results of such servers.

Although it was noted earlier that the coiled coil prediction methods are sensitive for highly charged false positives,[28,52] it cannot be ruled out that certain CSAH motifs could transiently dimerize/multimerize into coiled

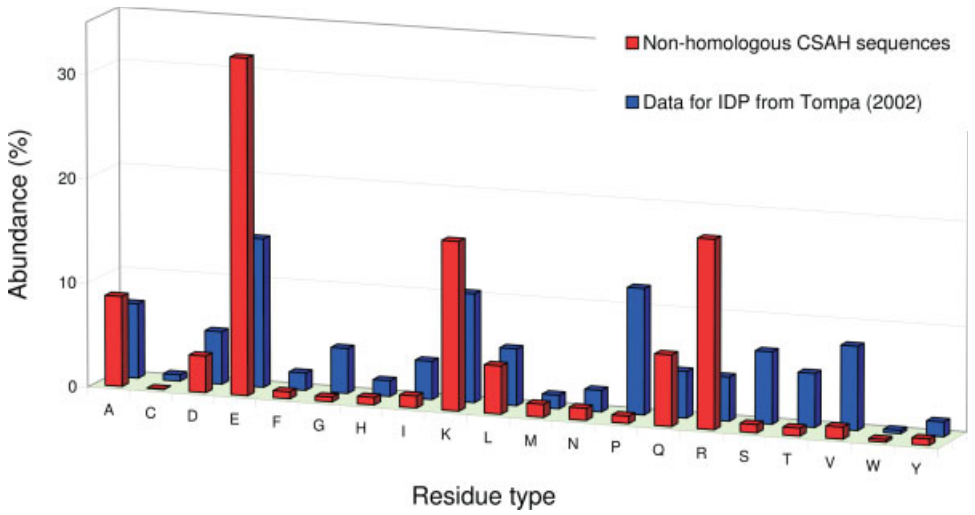

Figure 6

Average amino acid compositions of the detected CSAH sequences and intrinsically unstructured proteins. Data for the homology-filtered CSAH proteins and those reported by Tompa[51] are shown. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

coil like structures *in vivo* similar to that suggested for myosin 6.[34] CSAH dimers are unlikely to form canonical knobs-into-holes structures, but they are rather held together by electrostatic interactions between the chains staggered by the charge frequencies [as shown in Fig. 5(A)]. The fact that we found conventional coiled coil regions instead of single α-helices in several orthologs of predicted CSAH containing proteins (e.g., GLE1, M4K4, and NONO) makes it plausible that the CSAH motifs could be involved in transient interactions with partner molecules. The detected CSAH motifs are distinct sequences that could be, at least in theory, recognized more or less specifically by partner proteins. Thus, similarly to intrinsically disordered regions where, despite the seemingly similar structural features, different proteins have well-defined biological functions, we propose the existence of distinct roles and/or partners for each CSAH motif. CSAH motifs could form relatively rigid or flexible rods for a variety of purposes or could be dynamic structures capable of forming other structures [π-helix formation was observed in a molecular dynamics simulation of the caldesmon charged helix[15]] or undergo substantial folding-unfolding transitions. Sequences with ambiguously predicted secondary structure can act as conformational switches due to their versatile nature, a concept first introduced by Young *et al.*[53]

In our search for CSAHs in PDB no protein was detected by both methods. This result, although indirectly, reinforces the dynamic nature of the CSAH motif on the one hand, and the necessity of using the consensus of different approaches in CSAH prediction on the other. Note that the SGD database (http://www.yeastgenome.org) contains a putative homolog of the repeat region of MNN4 with known structure (PDB ID 1W9R, the r2 domain of the cholin binding protein a, the major adhesin of *S. pneumoniae*), a three-helix bundle protein, for which, however, neither FT_CHARGE nor SCAN4CSAH identifies CSAH segments. We believe that this protein has incidentally a similar but characteristically different charge pattern and it is not a true homolog of MNN4. It should also be noted that the highly charged and repetitive nature of CSAH motifs would make them difficult to study by atomic resolution structure determination methods.

We have searched for relatively long CSAHs; however, it can not be ruled out that shorter peptide segments in certain proteins also form stable α-helices stabilized by salt bridge networks or by other means. For instance, Liu *et al.*[54] identified a 26-residue fragment in human programmed cell death 5 protein that was shown by NMR spectroscopy to form a stable α-helix, however, with smaller charge density (not detected by our methods). This finding calls for caution in evaluating the intramolecular interactions necessary for helix stabilization and raises the possibility that single α-helices might be even more widespread than we can currently infer from our detection methods. Ribosomal protein L9 was also shown to contain a 17 residue CSAH[4] that was undetected by our FT-CHARGE method because of its shortness. [It should also be noted that the known CSAH domain of myosin 10 could be detected by FT_CHARGE only when segment lengths shorter than 64 residues were used.]

Although our results and literature data,[3–6] including computer simulation[15] prove the existence of the CSAH

motif, the explanation of its structural stability is not straightforward. The exact role of ionic interactions in helix stabilization is not yet clear. Using a systematic mutation approach for lysozyme, Takano *et al.*[55] showed that salt bridges near the surface exert only a slight stabilizing effect as the dielectric constant on the protein surface is greater than in the core, rendering ionic interactions weaker. However, although in CSAHs all charged residues are on the surface, their extreme density together with the regular periodicity of the opposite charges may contribute to stability to a greater extent than isolated salt bridges. Moreover, not only the ionic interactions can stabilize this structure: the long side chains could efficiently sequester water away from the vicinity of the backbone amides to promote intramolecular hydrogen bond formation.[12,13] Another possible explanation for the stability of CSAH motifs is the flexibility of the salt bridge network (i.e., the continuous breaking and reformation of ionic bonds with different partners), giving it an entropic advantage.

CONCLUSIONS

In this article we have shown that highly charged single α-helices are not rare exceptions in the protein world. Our consensus approach, using two conceptually different prediction algorithms for detecting such motifs yielded 69 consensus hits from an analysis of the Swiss-Prot database. This number is likely an underestimation of CSAH motifs since we used strict criteria for selecting candidate CSAH containing proteins and searched for only relatively long motifs. Three CSAH domains were chosen and shown experimentally to acquire α-helical structure in aqueous environment. We propose that these segments are functionally versatile, comprising a previously unrecognized general protein sequential and structural motif. Structural stabilization, dynamic behavior and functional significance of CSAH motifs merit further investigations.

ACKNOWLEDGMENTS

The authors thank Michel Espinoza-Fonseca and Gergely Zahoránszky for reading the manuscript and giving useful advices. The help of Balázs Szappanos with the coiled coil predictions is gratefully acknowledged. Z.G. was supported by a János Bolyai Postdoctoral Fellowship.

REFERENCES

1. Chou PY, Fasman GD. Prediction of protein conformation. Biochemistry 1974;13:222–245.
2. Robson B, Suzuki E. Conformational properties of amino acid residues in globular proteins. J Mol Biol 1976;107:327–356.
3. Knight PJ, Thirumurugan K, Xu Y, Wang F, Kalverda AP, Stafford WF, III, Sellers JR, Peckham M. The predicted coiled-coil domain of myosin 10 forms a novel elongated domain that lengthens the head. J Biol Chem 2005;280:34702–34708.
4. Kuhlman B, Yang HY, Boice JA, Fairman R, Raleigh DP. An exceptionally stable helix from the ribosomal protein L9: implications for protein folding and stability. J Mol Biol 1997;270:640–647.
5. Wang CL, Chalovich JM, Graceffa P, Lu RC, Mabuchi K, Stafford WF. A long helix from the central region of smooth muscle caldesmon. J Biol Chem 1991;266:13958–13963.
6. Wang E, Wang CL. (*i*, *i* + 4) Ion pairs stabilize helical peptides derived from smooth muscle caldesmon. Arch Biochem Biophys 1996;329:156–162.
7. Lyu PC, Gans PJ, Kallenbach NR. Energetic contribution of solvent-exposed ion pairs to α-helix structure. J Mol Biol 1992;223:343–350.
8. Lyu PC, Liff MI, Marky LA, Kallenbach NR. Side chain contributions to the stability of α-helical structure in peptides. Science 1990;250:669–673.
9. Lyu PC, Marky LA, Kallenbach NR. The role of ion pairs in α-helix stability: two new designed helical peptides. J Am Chem Soc 1989;111:2733–2734.
10. Marqusee S, Baldwin RL. Helix stabilization by Glu-Lys+ salt bridges in short peptides of de novo design. Proc Natl Acad Sci USA 1987;84:8898–8902.
11. Olson CA, Spek EJ, Shi Z, Vologodskii A, Kallenbach NR. Cooperative helix stabilization by complex Arg-Glu salt bridges. Proteins 2001;44:123–132.
12. Garcia AE, Sanbonmatsu KY. α-helical stabilization by side chain shielding of backbone hydrogen bonds. Proc Natl Acad Sci USA 2002;99:2782–2787.
13. Vila JA, Ripoll DR, Scheraga HA. Physical reasons for the unusual α-helix stabilization afforded by charged or neutral polar residues in alanine-rich peptides. Proc Natl Acad Sci USA 2000;97:13075–13079.
14. Andrew CD, Penel S, Jones GR, Doig AJ. Stabilizing nonpolar/polar side-chain interactions in the α-helix. Proteins 2001;45:449–455.
15. Shepherd CM, van der Spoel D, Vogel HJ. Molecular dynamics simulations of peptides from the central domain of smooth muscle caldesmon. J Biomol Struct Dyn 2004;21:555–566.
16. Scholtz JM, Qian H, Robbins VH, Baldwin RL. The energetics of ion-pair and hydrogen-bonding interactions in a helical peptide. Biochemistry 1993;32:9668–9676.
17. Stephenson A. evd: Extreme value distributions. R News 2002; 2/2:31–32.
18. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2008. http://www.r-project.org.
19. Sharma D, Issac B, Raghava GP, Ramaswamy R. Spectral repeat finder (SRF): identification of repetitive sequences using Fourier transformation. Bioinformatics 2004;20:1405–1412.
20. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 2003;31:365–370.
21. Johnson WC. Analyzing protein circular dichroism spectra for accurate secondary structures. Proteins 1999;35:307–312.
22. Lobley A, Whitmore L, Wallace BA. DICHROWEB: an interactive website for the analysis of protein secondary structure from circular dichroism spectra. Bioinformatics 2002;18:211–212.
23. Wolf E, Kim PS, Berger B. MultiCoil: a program for predicting two- and three-stranded coiled coils. Protein Sci 1997;6:1179–1189.
24. McDonnell AV, Jiang T, Keating AE, Berger B. Paircoil2: improved prediction of coiled coils from sequence. Bioinformatics 2006;22: 356–358.
25. Morii H, Takenawa T, Arisaka F, Shimizu T. Identification of kinesin neck region as a stable α-helical coiled coil and its thermodynamic characterization. Biochemistry 1997;36:1933–1942.
26. Delorenzi M, Speed T. An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. Bioinformatics 2002;18:617–625.
27. Lupas A. Prediction and analysis of coiled-coil structures. Methods Enzymol 1996;266:513–525.

28. Lupas A, Van Dyke M, Stock J. Predicting coiled coils from protein sequences. Science 1991;252:1162–1164.
29. Dosztanyi Z, Csizmok V, Tompa P, Simon I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 2005;21:3433–3434.
30. Dosztanyi Z, Csizmok V, Tompa P, Simon I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol 2005;347:827–839.
31. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A. Pfam: clans, web tools and services. Nucleic Acids Res 2006;34:D247–D251.
32. Brown JH, Yang Y, Reshetnikova L, Gourinath S, Suveges D, Kardos J, Hobor F, Reutzel R, Nyitray L, Cohen C. An unstable head-rod junction may promote folding into the compact off-state conformation of regulated myosins. J Mol Biol 2008;375:1434–1443.
33. Kumar S, Tsai CJ, Nussinov R. Factors enhancing protein thermostability. Protein Eng 2000;13:179–191.
34. Park H, Ramamurthy B, Travaglia M, Safer D, Chen LQ, Franzini-Armstrong C, Selvin PR, Sweeney HL. Full-length myosin VI dimerizes and moves processively along actin filaments upon monomer clustering. Mol Cell 2006;21:331–336.
35. Huber PA. Caldesmon. Int J Biochem Cell Biol 1997;29:1047–1051.
36. Wang CL, Chalovich JM, Graceffa P, Lu RC, Mabuchi K, Stafford WF. A long helix from the central region of smooth muscle caldesmon. J Biol Chem 1991;266:13958.
37. Mabuchi K, Wang CL. Electron microscopic studies of chicken gizzard caldesmon and its complex with calmodulin. J Muscle Res Cell Motil 1991;12:145–151.
38. Sohda M, Misumi Y, Yamamoto A, Yano A, Nakamura N, Ikehara Y. Identification and characterization of a novel Golgi protein, GCP60, that interacts with the integral membrane protein giantin. J Biol Chem 2001;276:45298–45306.
39. Sbodio JI, Hicks SW, Simon D, Machamer CE. GCP60 preferentially interacts with a caspase-generated golgin-160 fragment. J Biol Chem 2006;281:27924–27931.
40. Yao Z, Zhou G, Wang XS, Brown A, Diener K, Gan H, Tan TH. A novel human STE20-related protein kinase, HGK, that specifically activates the c-Jun N-terminal kinase signaling pathway. J Biol Chem 1999;274:2118–2125.
41. Kendirgi F, Rexer DJ, Alcazar-Roman AR, Onishko HM, Wente SR. Interaction between the shuttling mRNA export factor Gle1 and the nucleoporin hCG1: a conserved mechanism in the export of Hsp70 mRNA. Mol Biol Cell 2005;16:4304–4315.
42. Tusnady GE, Simon I. The HMMTOP transmembrane topology prediction server. Bioinformatics 2001;17:849–850.
43. Jigami Y, Odani T. Mannosylphosphate transfer to yeast mannan. Biochim Biophys Acta 1999;1426:335–345.
44. Sorensen HP, Hedegaard J, Sperling-Petersen HU, Mortensen KK. Remarkable conservation of translation initiation factors: IF1/eIF1A and IF2/eIF5B are universally distributed phylogenetic markers. IUBMB Life 2001;51:321–327.
45. Lee JH, Choi SK, Roll-Mecak A, Burley SK, Dever TE. Universal conservation in translation initiation revealed by human and archaeal homologs of bacterial translation initiation factor IF2. Proc Natl Acad Sci USA 1999;96:4342–4347.
46. Takai Y, Nakanishi H. Nectin and afadin: novel organizers of intercellular junctions. J Cell Sci 2003;116:17–27.
47. Straub T, Grue P, Uhse A, Lisby M, Knudsen BR, Tange TO, Westergaard O, Boege F. The RNA-splicing factor PSF/p54 controls DNA-topoisomerase I activity by a direct interaction. J Biol Chem 1998;273:26261–26264.
48. Giovannone B, Lee E, Laviola L, Giorgino F, Cleveland KA, Smith RJ. Two novel proteins that are linked to insulin-like growth factor (IGF-I) receptors by the Grb10 adapter and modulate IGF-I signaling. J Biol Chem 2003;278:31564–31573.
49. Lacroix E, Viguera AR, Serrano L. Elucidating the folding problem of alpha-helices: local motifs, long-range electrostatics, ionic-strength dependence and prediction of NMR parameters. J Mol Biol 1998;284:173–191.
50. Makowska J, Rodziewicz-Motowidlo S, Baginska K, Vila JA, Liwo A, Chmurzynski L, Scheraga HA. Polyproline II conformation is one of many local conformational states and is not an overall conformation of unfolded peptides and proteins. Proc Natl Acad Sci USA 2006;103:1744–1749.
51. Tompa P. Intrinsically unstructured proteins. Trends Biochem Sci 2002;27:527–533.
52. Gruber M, Soding J, Lupas AN. Comparative analysis of coiled-coil prediction methods. J Struct Biol 2006;155:140–145.
53. Young M, Kirshenbaum K, Dill KA, Highsmith S. Predicting conformational switches in proteins. Protein Sci 1999;8:1752–1764.
54. Liu D, Yao H, Chen Y, Feng Y, Chen Y, Wang J. The N-terminal 26-residue fragment of human programmed cell death 5 protein can form a stable α-helix having unique electrostatic potential character. Biochem J 2005;392:47–54.
55. Takano K, Tsuchimori K, Yamagata Y, Yutani K. Contribution of salt bridges near the surface of a protein to the conformational stability. Biochemistry 2000;39:12375–12381.