J. Mol. Biol. (1989) 206, 689–692

Are Coiled-coil Proteins Evolutionarily Related?

S. Subbiah

Committee on Higher Degrees in Biophysics
7, Divinity Avenue, Harvard University
Cambridge, MA 02138, U.S.A.

(Received 25 July 1988, and in revised form 22 December 1988)

A modification to the standard Needleman–Wunsch sequence comparison scheme is presented. In cases where high levels of sequence similarity may arise from a common structural motif, this method discriminates between common ancestry and similarity due to structural constraints alone. Use of this algorithm is illustrated with the coiled-coil motif in the cases of idealized coiled-coil sequences, intermediate filaments and reovirus hemagluttinin.

1. Introduction

Similarities of protein sequence are now commonly taken as indications of structural similarity and, by implication, of functional correspondence. The question arises, can detailed sequence analysis distinguish between sequence similarity arising from a common genetic origin and that arising from protein design constraints? That is, is the principal cause for an observed common motif rooted in common ancestry or common function? A good example is found in the α-helical coiled-coil structure (McLachlan *et al.*, 1975). Recent articles report a variety of newly determined sequences to be similar to those of coiled-coil proteins. Members of the coiled-coil family are largely structural proteins; tropomyosin, myosin rod, and intermediate-filament proteins such as keratin. The coiled-coil structure is characterized by a repeating pattern of hydrophobicity. This characteristic seven-residue repeating pattern, a-b-c-d-e-f-g, with positions a and d being primarily hydrophobic, permits the wrapping of two α-helices about each other and so allows the burying of their hydrophobic sides at the interface of the coil (McLachlan *et al.*, 1975; Doolittle *et al.*, 1978; Parry, 1981; McLachlan & Karn, 1982, 1983). I suggest here a test based on a modification of the standard Needleman–Wunsch (NW†) algorithm (Needleman & Wunsch, 1970) for determining the evolutionary relatedness between such coiled-coil sequences.

2. Methods

For any pair of sequences, the standard NW algorithm determines the optimal alignment and a measure of their similarity, k(orig). The significance of such an optimal alignment is commonly assessed as follows. First, each of the 2 sequences is scrambled N times to produce 2 sets of N randomized sequences. Then, one sequence from each set is selected and the similarity, $k(1)$, of this pair is computed. Similarly, this is repeated with further pairs of sequences drawn from the 2 sets. Thus a distribution of scores $k(1)$, $k(2)$, $k(4)\ldots\ldots k(N)$ is generated. The standard deviation, σ, of this distribution of scores and its average k(average) is used to calculate the significance score:

$$z(\text{regular}) = (k(\text{orig}) - k(\text{average}))/\sigma,$$

which has units of standard deviation. This score, z(regular), estimates the probability of the observed similarity score k(orig) occurring purely by chance. These similarity scores are computed using some particular scoring matrix that assigns a match score for every possible amino acid pair. Thus, for a given alignment, the individual match score at each position of the alignment is determined from the scoring matrix and these match scores are then totalled. Any relative insertions or deletions in the alignment are penalized by subtracting some pre-determined gap penalty score from the match score total to give the similarity score, k. It is customary to attempt a series of different gap penalties and select the gap penalty that simultaneously minimizes the number of relative gaps and maximizes the significance score, z(regular). For example, in practice a gap penalty between 1·5 and 3 is commonly found to be appropriate when using the scoring scheme of unity for identities and zero for non-identities. There are a number of scoring schemes available. Some of these are the UP matrix (Dayhoff, 1978), the SIM matrix (McLachlan, 1971), and the PAM matrix (Dayhoff, 1978). These 3 matrices are based on amino acid identity, on substitution frequency, and on mutation data, respectively. With each matrix there is an empirical cutoff value, z(cutoff), for the significance score that is used in measuring the level of relatedness of an alignment. From Doolittle's work using the UP matrix (Doolittle, 1981) it is clear that there are at least some instances where a significance score of up to 5·2 has been found in cases where it is unlikely that there

† Abbreviation used: NW, Needleman–Wunsch.

0022-2836/89/080689-04 $03.00/0

could be any ancestral relatedness. On the other hand, he also reported cases with significance scores as low as 2·45 where ancestral relatedness is highly likely. In many of these latter cases there are other biological grounds for expecting such relatedness. It seems clear, however, that when z(regular) is below 2·4 there is unlikely to be any case for ancestral relatedness, and when z(regular) is greater than 5·2 the case for ancestral relatedness is very strong. Within the range 2·4 to 5·2, the strength of the case for ancestral relatedness depends on whether other independent and supporting evidence exists. If such independent evidence does not exist, while the case for relatedness cannot be ruled out it is certainly not strong. Similarly, a range of 5 to 6 is considered the cutoff for relatedness with both the SIM matrix (McLachlan, 1971) and the PAM matrix (Dayhoff, 1978; Lipman & Pearson, 1985). In general, the values of these cutoffs are dependent on the sequence length. The values quoted refer to typical sequence lengths that are greater than 50 or so residues.

The modification I present here eliminates incidental high scores resulting from common structural motifs and involves structurally constrained randomization. The change occurs in the "scrambling" stage of the NW procedure. For the case of the coiled-coil, the regularly employed uniform shuffle is replaced by 2 uniform shuffles; one amongst the residues occupying the hydrophobic sites of all the heptads and the other amongst the remaining heptad residues. This "heptad shuffle" should be performed in cases where there is evidence for common coiled-coilness and where standard NW analysis results in high levels of sequence similarity (i.e. z(regular) > z(cutoff)). If the z value so attained, z(heptad), is less than z(cutoff), the sequences have no more similarity than that expected on the basis of structural similarity alone, and logic dictates that a common genetic origin is unlikely. If z(regular) > z(heptad) > z(cutoff), the case for common ancestry remains strong.

In order to conclude that a decrease in significance actually arises from the common heptad character of presumed coiled-coils, in practice certain further controls have to be performed. That is even for a pair of sequences lacking heptads, proceeding from a regular shuffle to a heptad shuffle will result in a drop in significance, owing to the finite sequence lengths and the limited number of randomization trials. This statistical drop, arising from the change in shuffling scheme alone, must be estimated and eliminated before concluding that the difference (z(regular) − z(heptad)) truly implies the presence of heptads. This estimate is obtained by jointly randomizing the original sequences in register and then calculating z(heptad) on this randomized pair. This new z(heptad) is in effect a control, z(control). If z(control) is found to be less than or equal to the original z(heptad), the overall drop in significance is probably an artifact of the approach and the candidate coiled-coil pattern is erroneous. By the same logic, if z(regular) > z(control) > z(heptad) > z(cutoff), the presence of the coiled-coil in consideration is confirmed.

Finally, a more correct way to perform a modified shuffle would be to shuffle each of the seven positions a, b, c, d, e, f and g separately instead of the 2-bin shuffle outlined above. This would indeed be the preferable way if the lengths of the sequences involved were very great. For such a 7-bin shuffle to be statistically meaningful, there should be significantly more than 20 residues in each of the 7 bins. This would imply that such a 7-bin approach should be applied only in cases where coiled-coil sequences are much longer than $20 \times 7 = 140$ residues. In general, the cases studied in this work do not have such long runs of heptads. Therefore, the simpler 2-bin approach has been used throughout.

3. Results

(a) *A model test case*

In order to test the discriminating ability of the scheme, random sequences with and without the coiled-coil motif were generated and subsequently subjected to the various shuffles. The results (Table 1) meet the following three criteria for successful discrimination. (1) Unrelated random sequences possessing heptads result in z(regular) > z(control) > z(cutoff) > z(heptad). (2) Related (in this case identical) coiled-coil sequences satisfy z(regular) > z(control) > z(heptad) > z(cutoff). (3) Related (in this case identical) sequences lacking coiled-coils produce the expected z(regular) > z(control) = z(heptad) > z(cutoff). Since the algorithm seemed promising, two biological applications of the scheme were considered.

(b) *Re-examining some well-established cases of proteins with heptad coiled-coils*

Many structural proteins like the keratins and the myosins have been shown to possess a coiled-coil motif (McLachlan *et al.*, 1975). Some of these proteins were examined to see whether the heptad shuffle protocol could distinguish between true relatedness and that arising from structural limitations. Two groups of three protein sequences each were considered. The first set consists of bovine keratin, *Xenopus* keratin and human nuclear lamin

Table 1

Similarity measures for random sequences

	R1/R2	R1/R1	R3/R3
z(regular)	3·667	62·449	65·836
z(control)	2·575	57·904	59·625
z(heptad)	0·311	39·487	59·759

Random sequences, each 100 residues long, were generated under a heptadic constraint of only hydrophobic residues at alternating intervals of 3 and 4 and hydrophilics elsewhere. (Real coiled-coils do deviate from this ideal, i.e. occasional misplaced residues, skip residues, etc.). Pairs of these "evolutionarily unrelated" random sequences, R1 and R2, were subjected to regular NW analysis, first using the general shuffle and then using the heptad shuffle. Subsequently, z(control) was estimated by jointly randomizing the original sequences in a regular shuffle and repeating NW analysis with the same heptad shuffle. The gap penalty was varied so as to maximize z and minimize the number of gaps associated with the optimal alignment. The required number, N, of random trials was estimated by varying it until the standard deviation σ of the distribution converged to the required accuracy. The testing was done using the UP matrix. Next, identical coiled-coil sequences, i.e. R1 and R1, were treated in the same way and z(regular), z(heptad) and z(control) obtained. Finally, the protocol was repeated with identical sequences that lack the coiled-coil pattern (i.e. R3 and R3).

Table 2

Similarity measures for 2 sets of well-known coiled-coil sequences

A. *Set 1*

Optimal gap penalty	Bovine keratin		*Xenopus* keratin		Nuclear lamin	
				24		24
z(regular)	Bovine keratin		19·927		5·071	
z(heptad)			18·421		5·201	
Optimal gap penalty				12		24
		3		↗		18
	Xenopus keratin		SIM matrix results		5·332	
	20·323				5·106	
		5				24
		10		10		
	Nuclear lamin		2·158		UP matrix results	
	3·459		1·789			
		10		10	↙	

B. *Set 2*

Optimal gap penalty	Chicken desmin		Hamster vimentin		Neuro-filament	
				18		18
z(regular)	Chicken desmin		24·625		15·945	
z(heptad)			18·534		11·652	
Optimal gap penalty				12		12
		5	↗			24
	Hamster vimentin		SIM matrix results		16·013	
	44·819				11·773	
	33·889					
		5				12
		10		5		
	Neuro-filament		21·707		UP matrix results	
	21·794		14·664			
	15·624					
		10		10	↙	

The protocol of Table 1 was carried out on each of the following 2 sets of sequences. Set 1, *Xenopus laevis* embryonal cytokeratin (233 to 331), bovine type I cytokeratin (237 to 335) and human nuclear lamin (234 to 333). Set 2, chicken desmin (255 to 352), hamster vimentin (257 to 354) and porcine neurofilament (260 to 357). NW analysis was performed over all pairs of sequences within each set, with 200 randomizations per comparison. Each pair was first compared using the UP matrix with the gap penalty varying as 0, 1, 2, 3, 5 and 10. This was repeated with the SIM matrix employing the gap penalties 12, 18, 24 and 100. In each case, the gap penalty that simultaneously maximized the *z* value and minimized the number of gaps was chosen. With the chosen gap penalty, the NW analysis was repeated for an increased number of random trials until the standard deviation had converged. The inset key panel shows the layout of the results for both *z*(regular) and *z*(heptad). The upper right of each matrix of results in A and B contains the scores for comparisons based on the SIM matrix, while the lower left contains similar results with the UP matrix. It is clear from A that the case for ancestral relatedness between human nuclear lamin and both the keratins is not particularly strong. Not surprisingly, however, both keratins are clearly related to each other. From B it is clear that all 3, desmin, vimentin and neurofilament, are strongly related to each other, over and beyond their structural similarity.

(Bader *et al.*, 1986; McKeon *et al.*, 1986). It is clear from Table 2A that the two keratins are, as to be expected, truly related. The relation between nuclear lamin and the keratins, on the other hand, is not as strong. The second set consists of chicken desmin, hamster vimentin and porcine neurofilament protein (Geisler *et al.*, 1984). In this case, the ancestral relatedness, over and beyond relatedness arising from similar structural constraints, of all three proteins is very convincing (Table 3B). Further, in both sets *z*(control) was found to be intermediate between the respective *z*(regular) and *z*(heptad) values.

(c) *The case of reovirus hemagluttinin*

The sequence of reovirus hemagluttinin has been reported to possess both the coiled-coil motif (residues 27 to 100) and homology suggesting common ancestry with rabbit skeletal tropomyosin and nematode myosin (residues 1391 to 1464: Basel-Duby *et al.*, 1985). There is no biological evidence that supports ancestral relatedness between either of the myosins and reovirus hemagluttinin. The application of the heptad shuffle protocol in conjunction with the SIM matrix shows that common ancestry is undetectable or distant in either case (Table 3B). Similarly, with the UP matrix there appears to be no ancestral relationship between rabbit tropomyosin and reovirus hemagluttinin (Table 3A). In the case of the nematode myosin and reovirus hemagluttinin pair, use of the UP matrix results in a significance score of 4·148. Although in statistical terms this is a reasonably strong indicator of relatedness, in the absence of additional supporting evidence and the low score obtained with the SIM matrix, the original claim of relatedness, which was made on the basis of a clearly strong significance score of 6·76, is possibly unwarranted.

4. Discussion

In conclusion, a method exists that is able to discriminate true sequence homology from "coincidental" homology owing to common structural constraints. Additionally, this protocol, under suitable circumstances, can determine the presence of such a structural motif above the "noise" of the sequence composition. In particular, the coiled-coil motif was used to illustrate this procedure. This procedure was tested on both model cases and on well-known coiled-coil proteins. It was then used to clarify a recently published result pertaining to the issue of common ancestry *versus* common goal. In this case of the reovirus hemagluttinin, the claim of ancestral relatedness to both rabbit tropomyosin and nematode myosin can be shown to be less

Table 3

Similarity measures for reovirus hemagluttinin and myosins

	A. *UP matrix*		B. *SIM matrix*	
	Tropomyosin	Myosin	Tropomyosin	Myosin
z(regular)	6·073	6·760	6·770	8·297
z(control)	3·427	5·237	5·031	6·941
z(heptad)	2·343	4·148	1·742	4·353

The protocol of Table 2 was carried out on the set reovirus hemagluttinin, nematode myosin and rabbit tropomyosin. The 1st element of the first row in A is the z(regular) score obtained when reovirus hemagluttinin was compared with rabbit tropomyosin using the UP matrix. The 2nd element of the first row in A shows z(regular) when the hemagluttinin–nematode myosin pair was considered. Similarly, the 2nd and 3rd rows in A show the z(control) and z(heptad) scores for the same 2 pairs of sequences. B differs from A in that the results were obtained with the SIM matrix rather than the UP matrix. In all cases, z(heptad) was computed using the coiled-coil assignment for reovirus hemagluttinin as the basis of the heptad shuffle z(control) was determined as outlined in Table 1. In every case, z(regular) > z(control) > z(heptad).

convincing after the protocol suggested here is employed. This approach could be used in general to eliminate other kinds of structural homology, for instance primary structural constraints in transmembrane domains in proteins, if a way can be found to codify the particular structural pattern in terms of protein sequence.

The author thanks S. C. Harrison, D. C. Wiley, A. Turkewitz, A. Mondragon and A. Aggarwal for their interest and helpful comments. Support is acknowledged from NSF grant no. CHE85 09574 (to S. C. Harrison, M. Karplus and D. C. Wiley).

References

Bader, B. L., Magin, T. M., Hatzfeld, M. & Franke, W. W. (1986). *EMBO J.* **5**, 1865–1875.

Basel-Duby, R., Jayasuriya, A., Chatterjee, D., Sonenberg, N., Maizel, J. V. & Fields, B. N. (1985). *Nature (London)*, **315**, 421–423.

Dayhoff, M. (1978). Editor of *Atlas of Protein Sequence and Structure*, vol. 5, National Biomedical Research Foundation, Silver Spring, MD.

Doolittle, R. F. (1981). *Science*, **214**, 149–159.

Doolittle, R. F., Goldbaum, D. M. & Doolittle, L. R. (1978). *J. Mol. Biol.* **121**, 311–325.

Geisler, N., Fischer, S., Vandekerckhove, J., Plessmann, U. & Weber, K. (1984). *EMBO J.* **3**, 2701–2706.

Lipman, D. J. & Pearson, W. R. (1985). *Science*, **227**, 1435–1441.

McKeon, F. D., Kirschner, M. W. & Caput, D. (1986). *Nature (London)*, **319**, 463–468.

McLachlan, A. D. (1971). *J. Mol. Biol.* **61**, 409–424.

McLachlan, A. D. & Karn, J. (1982). *Nature (London)*, **299**, 226–231.

McLachlan, A. D. & Karn, J. (1983). *J. Mol. Biol.* **164**, 605–626.

McLachlan, A. D., Stewart, M. & Smillie, L. B. (1975). *J. Mol. Biol.* **98**, 281–291.

Needleman, S. & Wunsch, C. (1970). *J. Mol. Biol.* **48**, 443–453.

Parry, D. A. D. (1981). *J. Mol. Biol.* **153**, 459–464.

Edited by B. W. Matthews