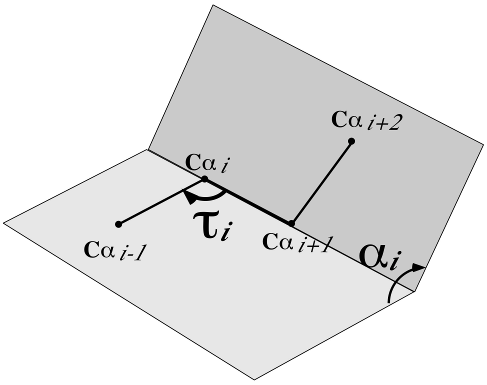
Supplementary Materials

I ALPHA ANGLES



Supplementary Fig. 1: The α angle is the angle made by two successive α-carbons.

II ALIGNMENT MATRIX

a)

		A	D	Q	R	T	A	L	M	**Q**	**K**	**T**	**A**
	0	0	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	0	0	0	4	0	0	0	0	0	4
D	0	0	0	0	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	1	0	0	0	0	**5**	1	0	0
R	0	0	0	0	0	0	0	0	0	1	**7**	0	0
T	0	0	0	0	0	0	0	0	0	0	0	**12**	4
A	0	0	0	0	0	0	0	0	0	0	0	4	**16**
L	0	0	0	0	0	0	0	0	2	0	0	0	8
M	0	0	0	0	0	0	0	0	0	2	0	0	0
Q	0	0	0	0	0	0	0	0	0	0	3	0	0
K	0	0	0	0	0	0	0	0	0	0	0	2	0
T	0	0	0	0	0	0	0	0	0	0	0	0	2
A	0	0	0	0	0	0	0	0	0	0	0	0	0

b)

		81	65	47	52	50	49	57	48	49	50	47	42	46	40	46
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
81	0	0	**11**	0	0	0	0	0	0	0	0	0	0	0	0	0
65	0	0	0	**7**	0	0	0	11	0	0	0	0	0	0	0	0
47	0	0	0	0	**14**	9	10	4	23	11	9	12	9	12	7	11
52	0	0	0	0	0	**25**	18	18	12	32	21	16	15	15	13	13
50	0	0	0	0	0	0	**35**	24	27	22	43	30	20	23	17	20
49	0	0	0	0	0	0	0	**40**	34	39	33	54	36	30	27	26
57	0	0	0	0	0	0	0	0	**44**	39	44	36	57	40	33	32
48	0	0	0	0	0	0	0	0	0	**55**	48	56	44	69	46	44
49	0	0	0	0	0	0	0	0	0	0	**66**	59	62	54	73	55
50	0	0	0	0	0	0	0	0	0	0	0	**75**	63	71	57	81
47	0	0	0	0	0	0	0	0	0	0	0	0	**83**	75	78	68
42	0	0	0	0	0	0	0	0	0	0	0	0	0	**94**	94	90
46	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**103**	107
40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**115**
46	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Supplementary Fig. 2: Examples of dynamic programming matrices using Swelfe. We compute only the upper half of the matrix. The diagonal is null. The best alignment is written in bold. In reality this matrix is not kept in memory as the SIM algorithm is used (Huang and Miller, 1991; Huang, et al., 1990).

a) Alignment of sequence data.

b) Alignment of structures coded as a string of α-angles (scores are decimal numbers but were truncated to fine down the scheme).

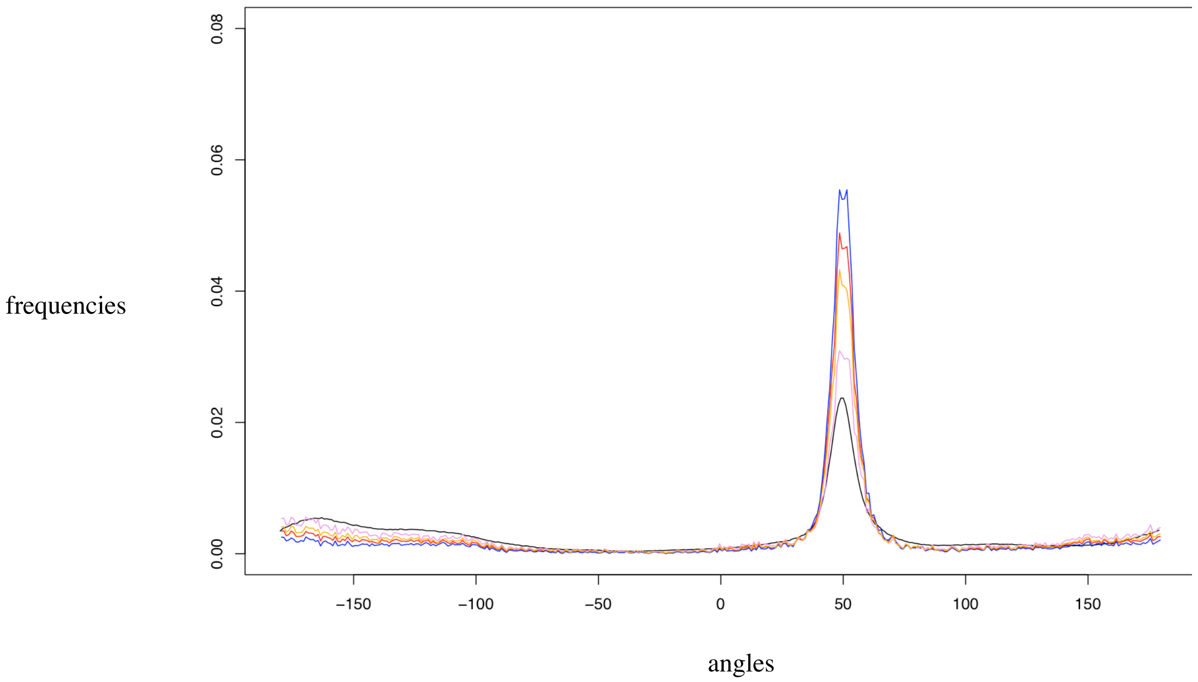
III SCORES

The scores used in the SIM alignment were adapted at each level. The nucleic acids substitution scores take into account the frequency of each nucleotide to produce more meaningful alignments (Achaz, et al., 2007). For amino acids sequence alignments we allow the use of any of the standard BLOSUM or PAM matrices. The empirical structural score for two matching α angles increases when the circular difference between them decreases and also accounts for the fact that some angles are much more frequent than others (Supplementary Fig. 3). This results in very frequent angles, e.g. originating from α-helices or β-sheets, having a lower score.

The default gap opening and extension parameters were set after analysis of an extensive range of values (Supplementary Table 1). For nucleic acid sequences, penalty for a gap opening corresponds to about 4 identities. For amino acid sequences it corresponds to between 1 and 2 identitie(s). For structures it corresponds to about 6 or 7 identities. Structural opening gaps are therefore more costly but if we decrease this penalty, we rapidly increase structural alignments with high RRMSD.

Supplementary Table 1 : Default scores used at each level ($S_{i,j}$). pi, pj : frequencies of nucleic acids for DNA sequences and normalized frequencies of α angles i and j in PDB for structures. Δangle is the angular difference between angles i and j (from 0 to 180°). The β parameter allows to give more or less weighting to angle frequencies p_i and p_j.

		Gap	
	Matching score	opening	extension
DNA sequence	$S_{i,j} = 0.5 \times \sigma_{(i,j)} \times \log_4(p_i p_j)$ $\sigma_{(i,j)} = 1$ if $i \neq j$; $\sigma_{(i,j)} = -1$ if $i = j$	-4	-1
Amino acid sequence	BLOSUM or PAM matrix	-8	-3
Three dimensional structure	$S_{i,j} = 30*[(1\text{-}p_i)(1\text{-}p_j)(1\text{-}\beta)+ \beta] - \mid \Delta\text{angle} \mid$ $\beta = 0.4$	-200 °	-50 °



Supplementary Fig. 3: frequencies of alpha angles in PDB structures (black) and in repeats found with β parameter = 0.4 (orange), β parameter = 0.5 (red), β parameter = 0.7 (blue), β parameter = 0.3 (violet). 500 best hits are kept for each β parameter values.

IV WATERMAN AND VINGRON METHOD TO ASSESS STATISTICAL SIGNIFICANCE

The Waterman and Vingron method (Waterman and Vingron, 1994) was used to calculate the significance of the repeats identified by dynamic programming. The general idea of the method is to generate a large number of random sequences (see Supplementary Fig. 4) with same length and composition as the gene (codons) or the protein (amino-acids) of interest. For each random sequence we identify the best local alignment score with the same algorithm as for the real sequence. The probability to get a score better than *t* in such a random sequence is then given by

$$P(P > t) = e^{-\gamma n^2 p^t}$$

where n is the length of the sequence. The parameters p and γ are estimated by a weighted linear regression:

$$\log(-\log(P(P > t))) = \log(\gamma n^2) + t\log(p)$$

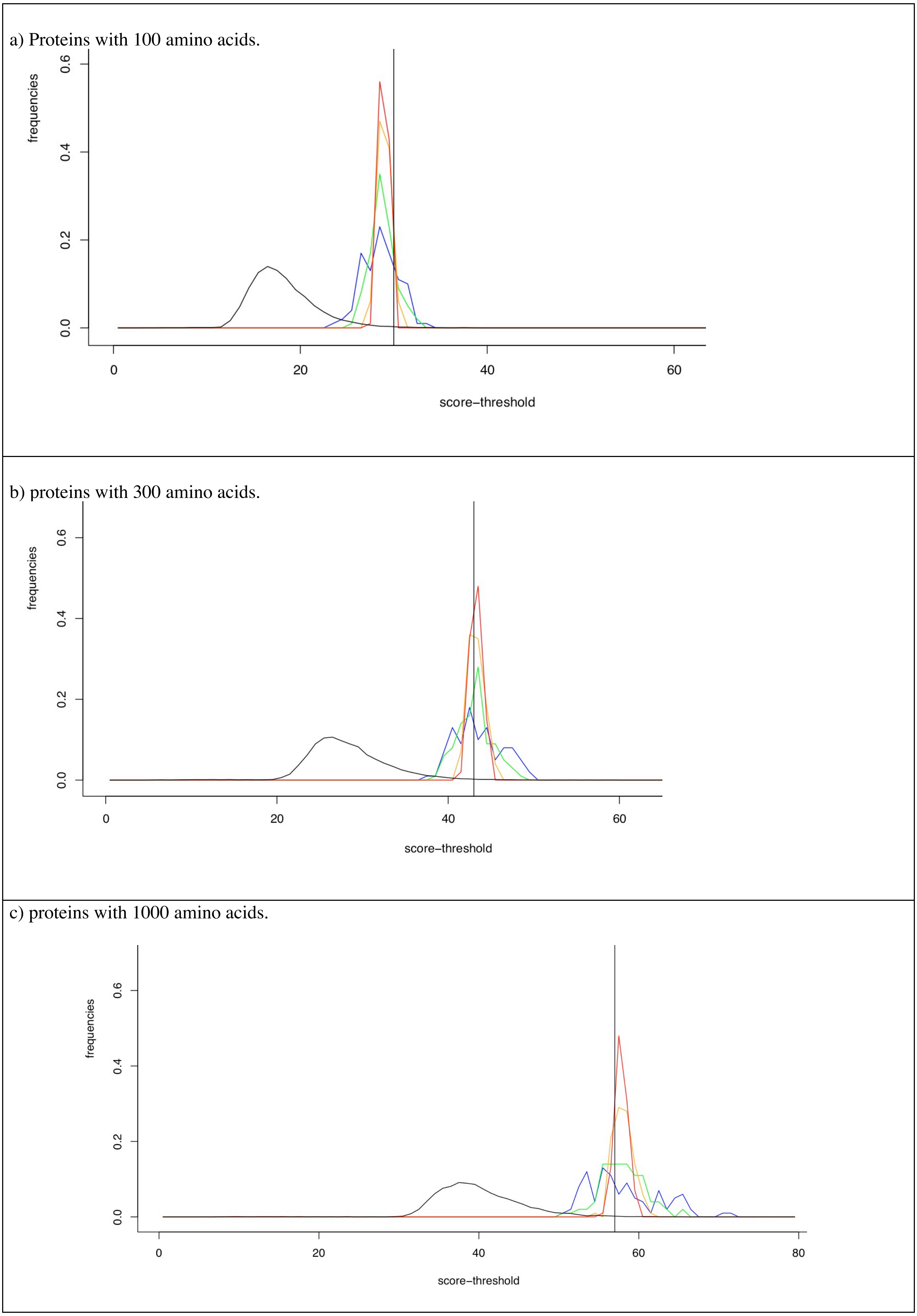
The use of weighted regression is motivated by the observation that random sequences may have sometimes the same score and this should be accounted for. The length of the sequence is corrected by subtracting the size of the mean length of local alignments for random sequences because sequences are finite (Mott, 2000).

We also implemented the "declumping estimation" method. In this case, we generate fewer random sequences (default 20) with the method described above. Then we calculate ~50 repeats for each random sequence with the declumping algorithm (successive repeats are independent), and we obtain about 1000 scores. In this case $\gamma n^2 p^t$ can be estimated by the average number of scores that exceed a threshold (E). The parameters p and γ are estimated by a weighted linear regression :

$$\log(E) = \log(\gamma n^2) + t\log(p).$$

Number of random sequences for statistics, first method.

We generate sequences of 100 AA, 300 AA and 1000 AA with the same composition in amino acids as the average composition of proteins in Uniprot. We make four different sets of experiments with 50 (blue curve in Fig. 4), 100 (green curve), 500 (orange curve) and 1000 (red curve) random sequences with same length and composition. With Swelfe, we define for each random sequence the score-threshold for a p-value of 0.01 with the Waterman-Vingron method. In the figure we represent the distribution of the thresholds for each set of experiments. We also compute the score distribution on an experiment with 10 000 random sequences (black curve). The vertical line represents the observed 99% threshold for the cumulative distribution of scores. Therefore, threshold values close to the vertical line are the most accurate. We conclude that 100 random sequences are sufficient to produce a threshold that is quite close to the real one while requiring limited computed time. Results were similar for DNA sequences (data not shown).



a) Proteins with 100 amino acids.

b) proteins with 300 amino acids.

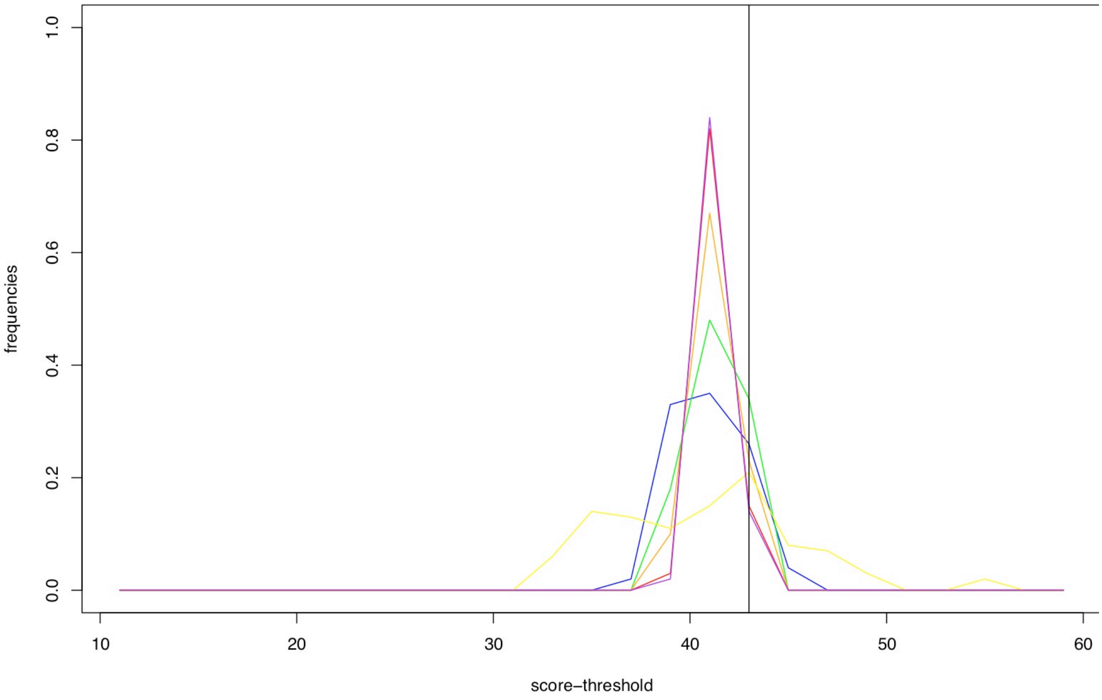
c) proteins with 1000 amino acids.

Supplementary Fig. 4 : Score-threshold obtained on 50 (blue), 100 (green), 500 (orange) and 1000 (red) random sequences of the same length and composition as the benchmark sequence for a p-value of 0.01.

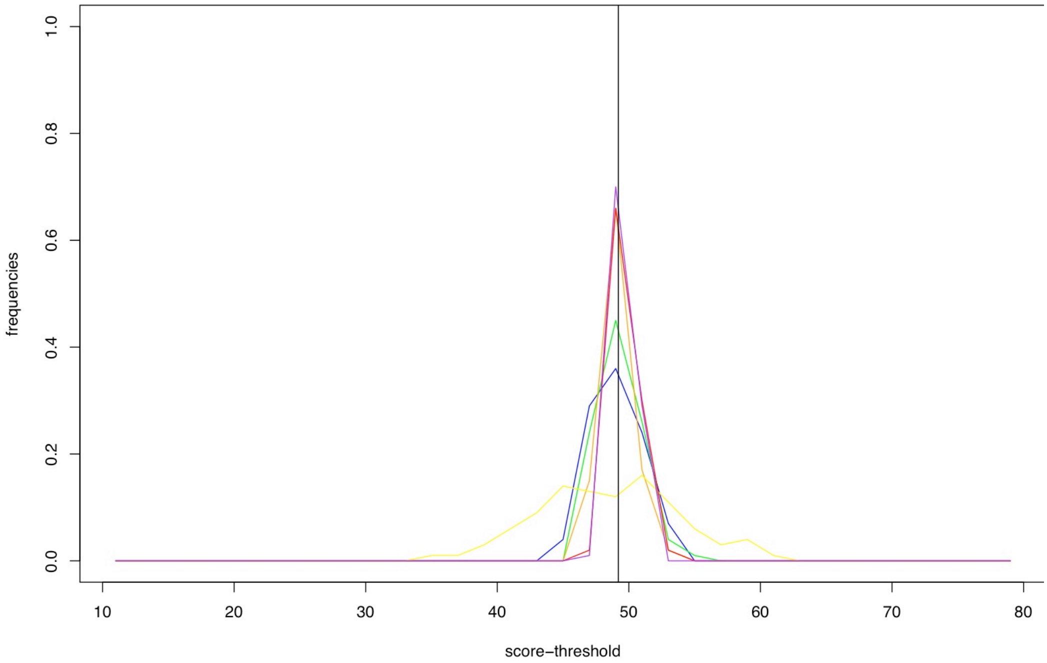
Supplementary figure 5 : number of random sequences for statistics, declumping estimation method.

The simulations are the same as above but with the declumping estimation method. We find that 20 random sequences provide a reasonable approximation. Results were similar for other sequence length ranges (data not shown). The threshold is slightly underestimated by the declumping estimation method in proteins.

a) protein sequences (300 AA)



b) nucleic acid sequences (1000 nt)



Supplementary Fig. 5 : Score-threshold obtained on 1 (yellow), 10 (blue), 20 (green), 30 (orange), 50(red) and 80 (violet) random sequences with same length and composition as the benchmark sequence for a p-value of 0.01.

V COMPARAISON WITH DALI

Supplementary Table 2 - Analysis of similarities between protein structures lacking secondary structure elements in (Novotny, et al., 2004) when using Swelfe (A) and DALI (B). We only indicate the results for the structures that were available on the DALI server. The numbers indicate the rank of the hit between the structures in the analysis of the structures for similarities against the reference database. NS means a non-significant hit.

A-Swelfe	**Targets**			
query	1b2i	1cea	1pml	5hpg
1b2i	1	8	14	6
1cea	10	1	8	3
1pml	NS	9	1	10
5hpg	12	2	11	1

B-DALI	**Targets**			
query	1b2i	1cea	1pml	5hpg
1b2i	1	9	38	20
1cea	30	1	33	NS
1pml	39	16	1	9
5hpg	39	4	38	1

Supplementary Table 3 – Analysis of similarities between protein structures regarded as difficult by (Novotny, et al., 2004). The table indicates the score of the significant matches using Swelfe and DALI.

		SWELFE	**DALI**
query	**target**	**rank**	**rank**
1bgeB	2gmfA	NS	43
3hlaB	2rheA	NS	615
2azaA	1pazA	NS	331
1cewI	1molA	30	35
1fxiA	1ubqA	215	NS
1cidA	2rheA	199	NS
1crlA	1edeA	351	271
1tenA	3hhrB	92	NS
1tieA	4fgfA	NS	177
2simA	1nsbA	683	74
1g61A	1jdwA	508	21

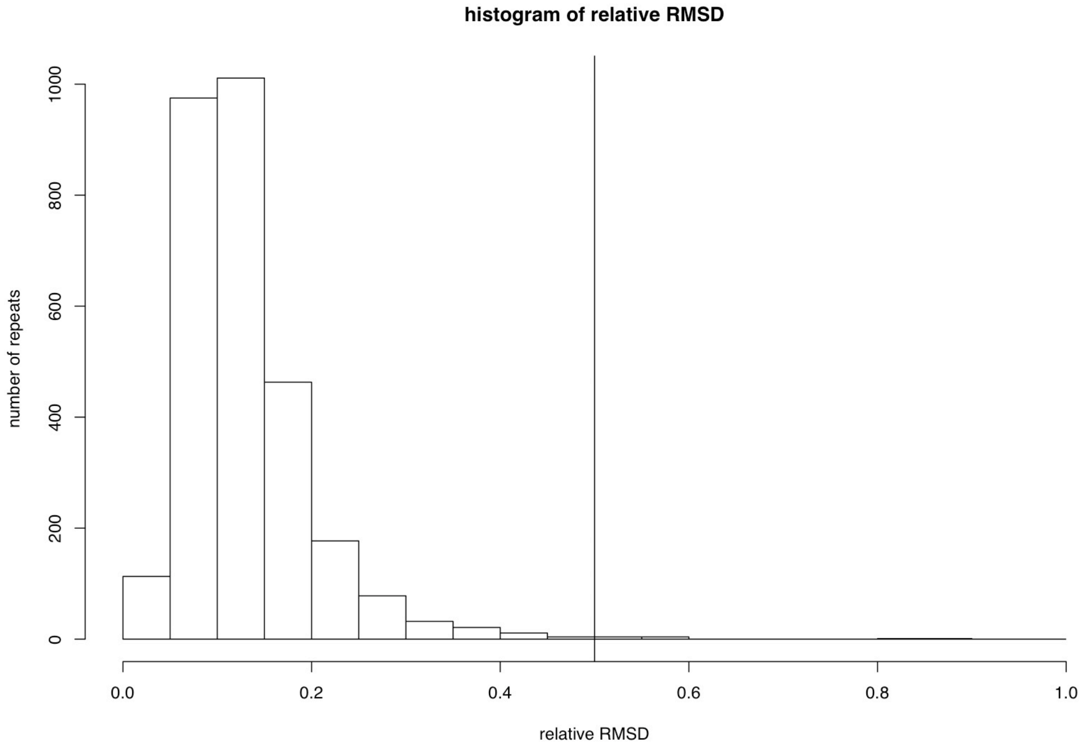
Supplementary Table 4 – Analysis of similarities between the structures of cyclophilins used by (Novotny, et al., 2004) when using Swelfe (A) and DALI (B). The numbers indicate the rank of the hit between the structures in the analysis of the structures for similarities against the reference database.

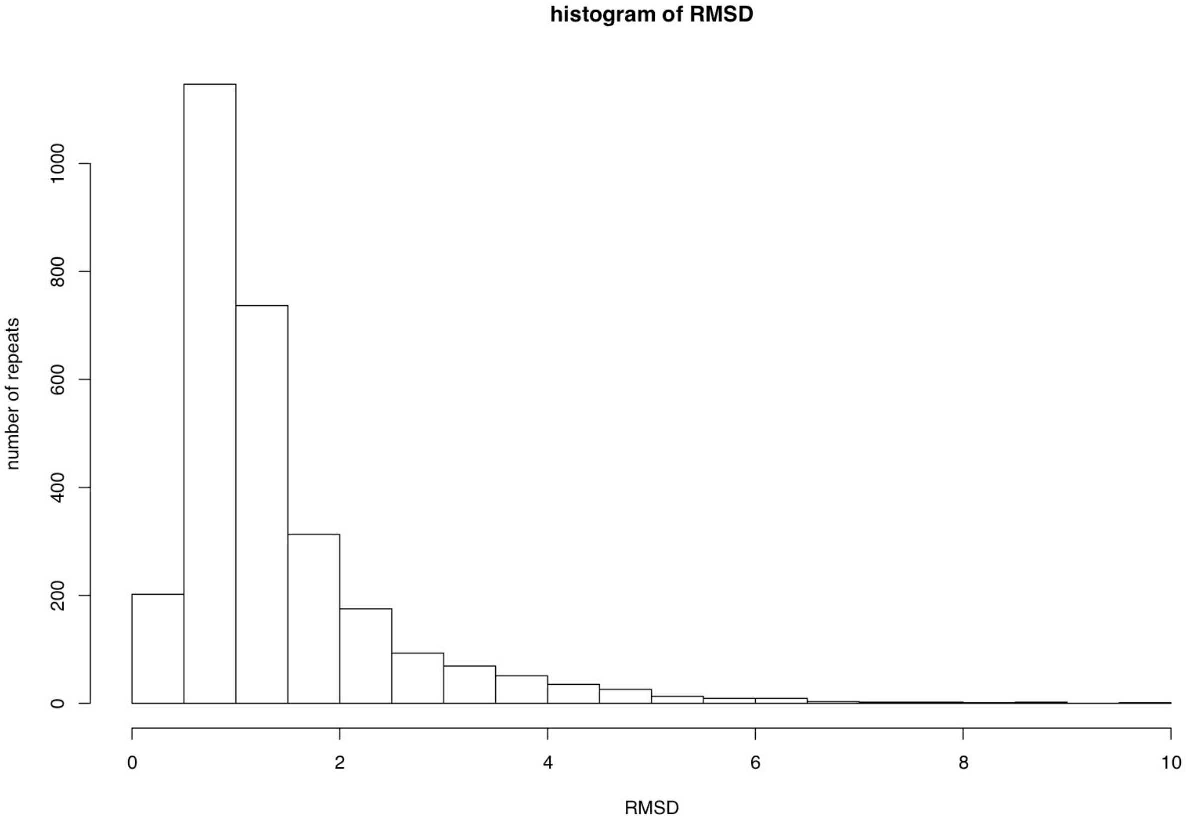
A-SWELFE	**Targets**							
query	1awq	1cyn	1qoi	1lop	1qng	2rmc	1dyw	1ihg
1awq	1	11	23	33	10	9	6	21
1cyn	15	1	23	33	17	2	13	19
1qoi	18	15	1	34	5	21	13	3
1lop	14	12	29	1	27	11	31	19
1qng	15	18	20	34	1	17	2	8
2rmc	14	3	23	33	17	1	12	19
1dyw	13	16	23	34	2	15	1	12
1ihg	24	16	11	44	6	18	5	1

	Targets							
query	1awq	1cyn	1qoi	1lop	1qng	2rmc	1dyw	1ihg
1awq	1	114	111	140	96	126	87	121
1cyn	117	1	103	139	104	4	101	105
1qoi	106	96	1	137	5	120	98	87
1lop	126	19	31	1	37	30	133	119
1qng	85	119	109	140	1	124	7	107
2rmc	114	5	107	139	110	1	103	115
1dyw	77	117	119	140	7	123	1	107
1ihg	112	80	64	139	18	119	16	1

VI RRMSD

A) RRMSD of repeats





b) RMSD of repeats

Supplementary Fig.6 : a) : Best repeats were calculated on 9537 3D structures with a score threshold (250), then the Relative RMSD is calculated. The great majority of the repeats are below 0.5 which is the threshold for significant repeats (Betancourt and Skolnick, 2001). b) RMSD calculated after the RRMSD threshold of 0.5.

REFERENCES

Achaz, G., Boyer, F., Rocha, E.P., Viari, A. and Coissac, E. (2007) Repseek, a tool to retrieve approximate repeats from large DNA sequences, *Bioinformatics*, **23**, 119-121.

Betancourt, M.R. and Skolnick, J. (2001) Universal similarity measure for comparing protein structures, *Biopolymers*, **59**, 305-309.

Huang, X. and Miller, W. (1991) A time-Efficient, Linear-Spaced Local Similarity Algorithm, *Adv In Appl Math*, **12**, 337-357.

Huang, X.Q., Hardison, R.C. and Miller, W. (1990) A space-efficient algorithm for local similarities, *Comput Appl Biosci*, **6**, 373-381.

Mott, R. (2000) Accurate formula for P-values of gapped local sequence and profile alignments, *J Mol Biol*, **300**, 649-659.

Novotny, M., Madsen, D. and Kleywegt, G.J. (2004) Evaluation of protein fold comparison servers, *Proteins*, **54**, 260-270.

Waterman, M.S. and Vingron, M. (1994) Rapid and accurate estimates of statistical significance for sequence data base searches, *Proc Natl Acad Sci U S A*, **91**, 4625-4628.