Supporting Information for

Protein-fingerprint data mining of a designed α-helical peptide array

Kenji Usui, Kin-ya Tomizaki, and Hisakazu Mihara*

Department of Bioengineering and the COE21 program, Graduate School of Bioscience and Biotechnology, Tokyo Institute of Technology, Nagatsuta-cho 4259 B-40, Midori-ku, Yokohama 226-8501, Japan

(Author for correspondence, E-mail: hmihara@bio.titech.ac.jp, Tel: +81-45-924-5756, FAX: +81-45-924-5833)*

METHODS

Data treatment for color scale 'protein fingerprints'

Data for 'protein fingerprints (PFPs)' used here were manipulated according to the standard procedure reported previously.[1-4] The file format used here was portable-pixel-map (.ppm) format. Each grid position was first assigned as three whole numbers corresponding to RGB color-codes representing increment response (0, 0, 0) (full black, minimum increasing value) to (255, 0, 0) (red) to (255, 255, 0) (yellow, maximum increasing value), which corresponds to all the fluorescence change rates (I/I_0) divided into 511 levels. The numbers of the grid were saved as a comma-separated-value (.csv) file including the three (or four) lines of ppm setting at the top of the file. The file was then saved in the portable-pixel-map format by simply adding '.ppm' to the filename. This file was opened by a graphic viewer software, resized and saved in other formats such as bitmap file format (.bmp).

Data manipulation using Euclidean distance and hierarchical clustering analysis (HCA)

The measure used to determine the similarity between two PFPs obtained from different target proteins is Euclidean distance.[5, 6] This is a common measure when considering the distance between two vectors. Before the Euclidean distance analyses were performed, the PFP must be normalized. Similarity between the normalized PFP patterns is measured by Euclidean distance in multidimensional space defined by each PFP. These should be represented by color-coding (yellow for the highest similarity and black for the lowest) as described in the previous section. Additionally, the hierarchical cluster analysis among the normalized PFPs was conducted. Ward's clustering algorithm was used and the dendrogram was obtained with the analyses of Euclidean distances using the Excel Macro program.[7] The horizontal axis represents the distance among normalized PFPs (left for PFPs with the highest similarity and right for PFPs with the lowest similarity).

Data manipulation using principal components analysis (PCA)

Principal components analysis (PCA) is a dimension reduction technique, *i.e.* PCA reduces the number of variables (features) to a more manageable size. In this study, Varimax rotation algorithm was used, and the results of analyses were obtained by cgi script program on the web site.[8]

References

1 K. Usui, M. Takahashi, K.Nokihara and H. Mihara, *Mol. Divers.* 2004, **8**, 209.
2 M. Takahashi, K. Nokihara and H. Mihara, *Chem. Biol.* 2003, **10**, 53.
3 K. Usui, T. Ojima, M. Takahashi, K. Nokihara and H. Mihara, *Biopolymers* 2004, **76**, 129.
4 D. Wahler, F. Badalassi, P. Crotti and J. L. Reymond, *Chem. Eur. J.*, 2002, **8**, 3211.
5 J. P. Goddard and J. L. Reymond *J. Am. Chem. Soc.* 2004, **126**, 11116.
6 J. Grognux and J. L. Reymond, *ChemBioChem* 2004, **5**, 826.
7 http://aoki2.si.gunma-u.ac.jp/lecture/stats-by-excel/vba/html/clustan.html (Japanese).
8 http://aoki2.si.gunma-u.ac.jp/BlackBox/BlackBox.html (Japanese).

No.	Name	Sequence
1	L8K6	LKKLLKLLKKLLKL
2	L8K4E2	LKKLLELLKKLLEL
3	L8K3E3	LEELLKLLKKLLEL
4	L8K2E4	LEELLKLLEELLKL
5	L6A2K6	LKKLLKALKKLLKA
6	L4A4K6	LKKLAKALKKLAKA
7	L2A6K6	LKKAAKALKKAAKA
8	A8K6	AKKAAKAAKKAAKA
9	L6A2K4E2	LKKLLEALKKLLEA
10	L4A4K4E2	LKKLAEALKKLAEA
11	L2A6K4E2	LKKAAEALKKAAEA
12	A8K4E2	AKKAAEAAKKAAEA
13	L6A2K3E3	LEELLKALKKLLEA
14	L4A4K3E3	LEELAKALKKLAEA
15	L2A6K3E3	LEEAAKALKKAAEA
16	A8K3E3	AEEAAKAAKKAAEA
17	L6A2K2E4	LEELLKALEELLKA
18	L4A4K2E4	LEELAKALEELAKA
19	L2A6K2E4	LEEAAKALEEAAKA
20	A8K2E4	AEEAAKAAEEAAKA
21	F2L6K6	LKKLLKFLKKLLKF
22	F2L6K4E2	LKKLLEFLKKLLEF
23	F2L6K3E3	LEELLKFLKKLLEF
24	F2L6K2E4	LEELLKFLEELLKF
25	F4L4K6	LKKLFKFLKKLFKF
26	F4L4K4E2	LKKLFEFLKKLFEF
27	F4L4K3E3	LEELFKFLKKLFEF
28	F4L4K2E4	LEELFKFLEELFKF
29	L8K4S2	LKKLLSLLKKLLSL
30	L8K3S3	LSSLLKLLKKLLSL
31	L8K2S4	LSSLLKLLSSLLKL
32	L8S6	LSSLLSLLSSLLSL
33	L6A2K4S2	LKKLLSALKKLLSA
34	L4A4K4S2	LKKLASALKKLASA
35	L2A6K4S2	LKKAASALKKAASA
36	A8K4S2	AKKAASAAKKAASA
37	L6A2K3S3	LSSLLKALKKLLSA
38	L4A4K3S3	LSSLAKALKKLASA
39	L2A6K3S3	LSSAAKALKKAASA
40	A8K3S3	ASSAAKAAKKAASA
41	L6A2K2S4	LSSLLKALSSLLKA
42	L4A4K2S4	LSSLAKALSSLAKA
43	L2A6K2S4	LSSAAKALSSAAKA
44	A8K2S4	ASSAAKAASSAAKA
45	L6A2S6	LSSLLSALSSLLSA
46	L4A4S6	LSSLASALSSLASA
47	L2A6S6	LSSAASALSSAASA
48	A8S6	ASSAASAASSAASA
49	F2L6K4S2	LKKLLSFLKKLLSF
50	F2L6K3S3	LSSLLKFLKKLLSF
51	F2L6K2S4	LSSLLKFLSSLLKF
52	F2L6S6	LSSLLSFLSSLLSF
53	F4L4K4S2	LKKLFSFLKKLFSF
54	F4L4K3S3	LSSLFKFLKKLFSF
55	F4L4K2S4	LSSLFKFLSSLFKF
56	F4L4S6	LSSLFSFLSSLFSF
57	L8K4Q2	LKKLLQLLKKLLQL
58	L8K3Q3	LQQLLKLLKKLLQL
59	L8K2Q4	LQQLLKLLQQLLKL
60	L8Q6	LQQLLQLLQQLLQL
61	L6A2K4Q2	LKKLLQALKKLLQA
62	L4A4K4Q2	LKKLAQALKKLAQA
63	L2A6K4Q2	LKKAAQALKKAAQA
64	A8K4Q2	AKKAAQAAKKAAQA
65	L6A2K3Q3	LQQLLKALKKLLQA
66	L4A4K3Q3	LQQLAKALKKLAQA
67	L2A6K3Q3	LQQAAKALKKAAQA
68	A8K3Q3	AQQAAKAAKKAAQA
69	L6A2K2Q4	LQQLLKALQQLLKA
70	L4A4K2Q4	LQQLAKALQQLAKA
71	L2A6K2Q4	LQQAAKALQQAAKA
72	A8K2Q4	AQQAAKAAQQAAKA
73	L6A2Q6	LQQLLQALQQLLQA
74	L4A4Q6	LQQLAQALQQLAQA
75	L2A6Q6	LQQAAQALQQAAQA
76	A8Q6	AQQAAQAAQQAAQA
77	F2L6K4Q2	LKKLLQFLKKLLQF
78	F2L6K3Q3	LQQLLKFLKKLLQF
79	F2L6K2Q4	LQQLLKFLQQLLKF
80	F2L6Q6	LQQLLQFLQQLLQF
81	F4L4K4Q2	LKKLFQFLKKLFQF
82	F4L4K3Q3	LQQLFKFLKKLFQF
83	F4L4K2Q4	LQQLFKFLQQLFKF
84	F4L4Q6	LQQLFQFLQQLFQF
85	L8R6	LRRLLRLLRRLLRL
86	L8R4E2	LRRLLELLRRLLEL
87	L8R3E3	LEELLRLLRRLLEL
88	L8R2E4	LEELLRLLEELLRL
89	L6A2R6	LRRLLRALRRLLRA
90	L4A4R6	LRRLARALRRLARA
91	L2A6R6	LRRAARALRRAARA
92	A8R6	ARRAARAARRAARA
93	L6A2R4E2	LRRLLEALRRLLEA
94	L4A4R4E2	LRRLAEALRRLAEA
95	L2A6R4E2	LRRAAEALRRAAEA
96	A8R4E2	ARRAAEAARRAAEA
97	L6A2R3E3	LEELLRALRRLLEA
98	L4A4R3E3	LEELARALRRLAEA
99	L2A6R3E3	LEEAARALRRAAEA
100	A8R3E3	AEEAARAARRAAEA
101	L6A2R2E4	LEELLRALEELLRA
102	L4A4R2E4	LEELARALEELARA
103	L2A6R2E4	LEEAARALEEAARA
104	A8R2E4	AEEAARAAEEAARA
105	F2L6R6	LRRLLRFLRRLLRF
106	F2L6R4E2	LRRLLEFLRRLLEF
107	F2L6R3E3	LEELLRFLRRLLEF
108	F2L6R2E4	LEELLRFLEELLRF
109	F4L4R6	LRRLFRFLRRLFRF
110	F4L4R4E2	LRRLFEFLRRLFEF
111	F4L4R3E3	LEELFRFLRRLFEF
112	F4L4R2E4	LEELFRFLEELFRF

Fig. S1 Numbers, names and sequences of peptides in the □-helical peptide library.[1]

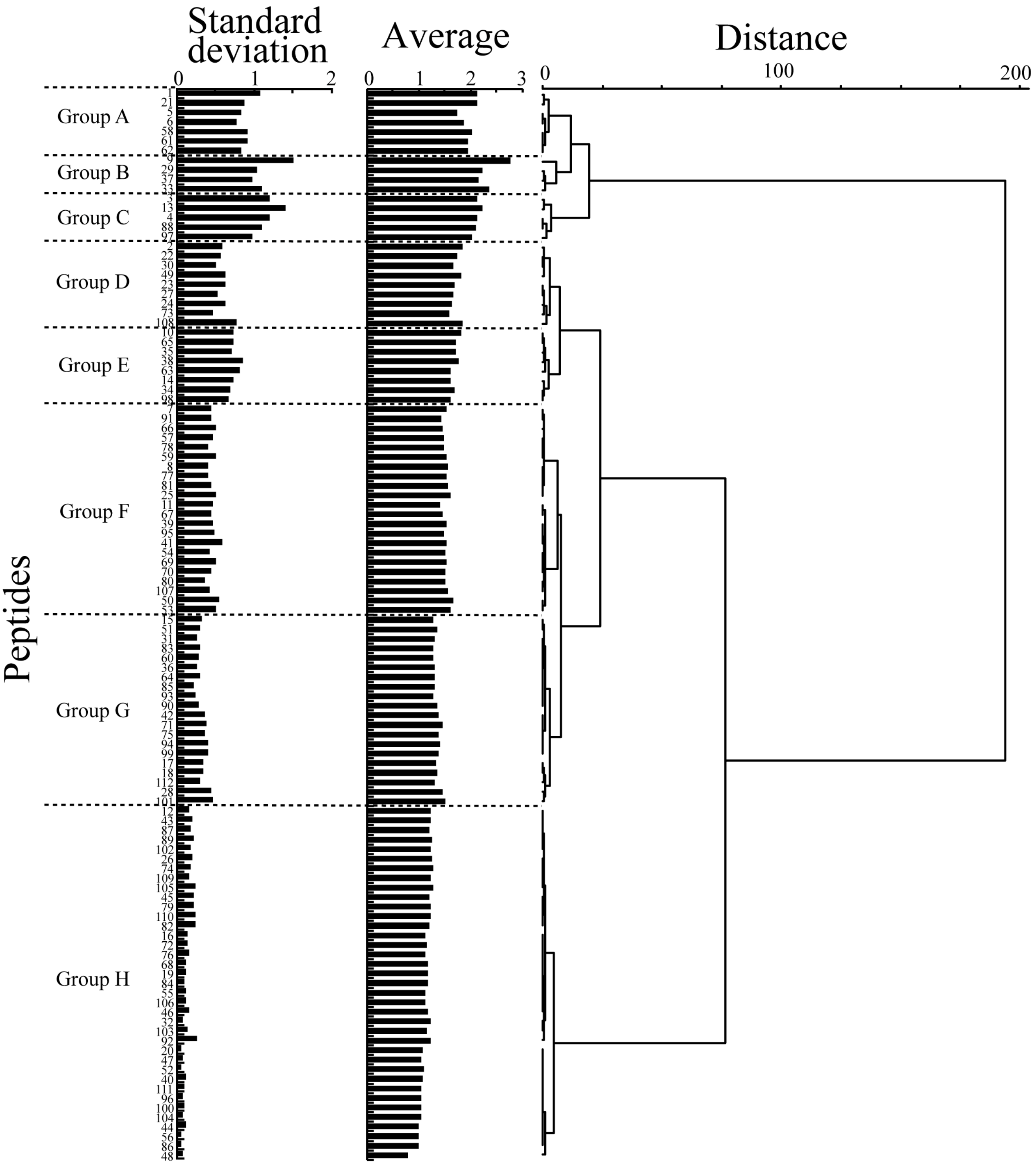

Fig. S2 Standard deviations (left) and averages (middle) of the PFP values of each peptide against the seven proteins [calmodulin (CaM), S-100 proteins (S-100), myosin, protein kinase A (PKA), □-lactoglobulin (□-LG), □-amylase, and insulin]. The clustering dendrogram of peptide divergences generated by the analysis of the Euclidean distances (right).