Data in Brief 9 (2016) 292–313

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Data from computational analysis of the peptide linkers in the MocR bacterial transcriptional regulators

Sebastiana Angelaccio [a], Teresa Milano [a], Angela Tramonti [b], Martino Luigi Di Salvo [a], Roberto Contestabile [a], Stefano Pascarella [a,*]

[a] Dipartimento di Scienze biochimiche "A. Rossi Fanelli", Sapienza Università di Roma, 00185 Roma, Italy
[b] Istituto di Biologia e Patologia Molecolari, Consiglio Nazionale delle Ricerche, 00185 Roma, Italy

ARTICLE INFO

Article history:
Received 18 July 2016
Received in revised form
24 August 2016
Accepted 30 August 2016
Available online 5 September 2016

Keywords:
Linker peptide
Linker length
MocR regulators
Linker engineering
PdxR
GabR
Hydrophobicity
Flexibility
Residue propensity
Dyad propensity

ABSTRACT

Detailed data from statistical analyses of the structural properties of the inter-domain linker peptides of the bacterial regulators of the family MocR are herein reported. MocR regulators are a recently discovered subfamily of bacterial regulators possessing an N-terminal domain, 60 residue long on average, folded as the winged-helix-turn-helix architecture responsible for DNA recognition and binding, and a large C-terminal domain (350 residue on average) that belongs to the fold type-I pyridoxal 5′-phosphate (PLP) dependent enzymes such aspartate aminotransferase. Data show the distribution of several structural characteristics of the linkers taken from bacterial species from five different phyla, namely Actinobacteria, Alpha-, Beta-, Gammaproteobacteria and Firmicutes.

Interpretation and discussion of reported data refer to the article "*Structural properties of the linkers connecting the N- and C-terminal domains in the MocR bacterial transcriptional regulators*"

DOI of original article: http://dx.doi.org/10.1016/j.biopen.2016.07.002
* Corresponding author. Fax: +39 06 49917566.
E-mail address: Stefano.Pascarella@uniroma1.it (S. Pascarella).

http://dx.doi.org/10.1016/j.dib.2016.08.064

(T. Milano, S. Angelaccio, A. Tramonti, M. L. Di Salvo, R. Contestabile, S. Pascarella, 2016) [1].

Specifications Table

Subject area	*Biology*
More specific subject area	*Structural properties of linkers in the bacterial transcriptional regulators*
Type of data	*Table, graph, figure*
How data was acquired	*Databank searches. Computational analysis*
Data format	*Raw, filtered, analyzed*
Experimental factors	*Analyses were mostly carried out with Perl, Python and R scripts and software for structural bioinformatics*
Experimental features	*Linker sequences were extracted from multiple sequence alignments of MocR regulators. Computational analysis defined the residue and residue dyads propensities and the distribution of physicochemical properties in the linker sequences.*
Data source location	*UniProt, RefSeq*
Data accessibility	*Data is within this article. Linker sequence sets are available at* https://sites.google.com/a/uniroma1.it/pascarellalab/home/resources

Value of the data

- Data represent the description of the structural properties of the peptide linkers connecting the N- and C-terminal domains in the MocR bacterial regulators.
- Data provide researchers with a framework to select specific MocR for experimental characterization.
- Data provide a support to design experiments for the investigation of properties of specific MocR: for example, experiments of site-directed mutagenesis, deletions or insertions of linker regions.
- Data can help interpretation of experimental data obtained from MocR studies.
- Data provide a framework to derive rules for *de-novo* design of peptide linkers with desired properties.

1. Data

Results derived from computational analysis of the inter-domain sequences of the peptide linker connecting the N-terminal and the C-terminal domain of the bacterial transcriptional regulators of the subfamily MocR are herein reported. Data are shown as tables describing linker statistics such as residue and dyad composition propensities, predicted secondary structure frequency, and box-plots showing the distribution of several structural properties. Moreover, plots of length distributions of linkers from two specific MocR subgroups, namely PdxR and GabR, are also reported.

Table 1
List of MocR regulators predicted to have linkers of length equal or greater to 60 residues.

UniProt code	Start[a]	End[b]	Length
A0A023C4T7_9PSED	88	148	60
A0A0B2AVS1_9ACTN	85	145	60
A0NP21_LABAI	80	140	60
I9W6R0_9RALS	87	147	60
W4CMK3_9BACL	121	181	60
A0A074LC92_PAEPO	82	143	61
I4N7I5_9PSED	87	148	61
A0A0D5NE20_9BACL	85	147	62
F8FPR4_PAEMK	106	168	62
G8QJ34_DECSP	85	147	62
V7DIJ8_9PSED	88	150	62
W4P2V0_9BURK	87	149	62
B9QZW6_LABAD	80	143	63
F3KUT6_9BURK	118	181	63
F7T5G0_9BURK	85	148	63
M2X958_9MICC	80	143	63
R9LS02_9BACL	83	146	63
S2WJB8_DELAC	89	152	63
A0A098SWK7_9PSED	88	152	64
A0A0J6J2M6_9PSED	88	152	64
A0A0F4KHT0_9ACTN	101	166	65
D5BN74_PUNMI	82	147	65
D7DQ74_METV0	90	156	66
K0YXF4_9ACTN	79	145	66
A0A077LFC1_9PSED	87	154	67
A0A095YU49_9FIRM	78	145	67
H0BWG7_9BURK	75	142	67
A0A087DUC1_9BIFI	78	146	68
A0A090ZGE9_PAEMA	83	152	69
A0A0A6Q9N6_9BURK	74	143	69
F3JEN8_PSESX	88	157	69
W0HH53_PSECI	88	157	69
A0A0A6QBJ9_9BURK	89	159	70
A0A0B4DLS5_9MICC	89	159	70
A0A088Y9M0_BURPE	88	159	71
A0A0F4JB47_9ACTN	62	135	73
A0A069DE36_9BACL	85	159	74
A0A087EGV8_9BIFI	105	181	76
A0A089I7M0_9BACL	82	158	76
A8SVX0_9FIRM	79	155	76
A0A089N895_9BACL	78	155	77
A0A0F5JX35_9BURK	84	161	77
A0A0E4CZM5_9BACL	90	168	78
A0A061LXN0_9MICO	84	163	79
A0A0A8BLT7_9BURK	89	168	79
R6HHE8_9ACTN	79	159	80
X4ZGS7_9BACL	84	164	80
A0A089HPN9_PAEDU	78	162	84
D2PX75_KRIFD	93	178	85
D3F8U9_CONWI	80	166	86
A0A0A4HID4_9PSED	88	179	91
F2RK57_STRVP	86	180	94
C7MPD0_CRYCD	79	174	95
F4QXL0_BREDI	83	179	96
A0A087AB73_9BIFI	78	175	97
A0A087E7D4_9BIFI	78	175	97
A0A0B4DPH0_KOCRH	85	183	98
F2RA50_STRVP	88	186	98
V6KRX5_STRRC	93	191	98
M8D4I1_9BACL	79	179	100

Table 1 (*continued*)

UniProt code	Start[a]	End[b]	Length
A0A0A3JRX6_BURPE	88	189	101
M8DED6_9BACL	80	183	103
A0A087BLK1_BIFLN	78	187	109
A0A087CXD8_9BIFI	78	187	109
S6CDU1_9ACTN	130	244	114
A0A0A6SYE7_9BURK	87	209	122
F5LR05_9BACL	82	209	127
A0A089IZ38_PAEDU	84	218	134
A0A0B6S8F7_BURGL	88	231	143
A0A087A119_9BIFI	78	222	144
A0A089MC10_9BACL	82	234	152
A0A089KZI8_9BACL	82	244	162

[a] Linker N-terminal sequence position.
[b] Linker C-terminal sequence position.

Table 2
Residue propensities in the linkers of length range 0–20.

	Actinobacteria		Alpha		Beta		Firmicutes		Gamma	
AA[a)]	P[b)]	Counts[c)]	P[b)]	Counts[c)]	P[b)]	AA[a)]	P[b)]	Counts[c)]	P[b)]	Counts[c)]
A	**1.23**	638	1.16	360	**1.24**	217	0.92	180	1.14	390
C	0.20	7	0.12	3	0.00	0	0.45	13	0.32	13
D	0.85	199	1.16	172	1.07	81	**1.30**	196	1.16	236
E	1.00	208	**1.37**	190	**1.44**	104	**1.40**	264	**1.35**	285
F	0.36	39	0.35	33	0.20	10	0.43	52	0.55	82
G	1.10	387	**1.20**	254	**1.43**	161	0.74	132	**1.21**	307
H	0.67	58	0.78	40	1.07	34	**1.60**	79	**1.49**	123
I	0.50	71	0.46	60	0.44	28	0.65	134	0.64	139
K	0.26	21	0.53	45	0.36	17	**1.26**	244	0.64	111
L	0.61	237	0.60	155	0.77	112	0.55	146	0.79	308
M	0.35	25	0.47	28	0.78	25	0.52	37	0.83	69
N	0.34	28	0.45	34	0.37	16	**1.68**	235	0.73	116
P	**2.90**	697	**2.69**	354	**2.13**	155	**1.71**	160	**1.80**	287
Q	0.73	86	1.12	97	**1.56**	90	1.16	118	**1.64**	276
R	**1.28**	389	**1.43**	257	1.02	103	0.93	110	0.86	174
S	1.01	234	1.17	182	1.02	89	**1.58**	283	**1.24**	312
T	1.18	288	0.96	135	1.12	85	**1.21**	188	0.94	191
V	0.79	257	0.66	117	0.71	70	0.69	125	0.90	216
W	0.41	25	0.28	10	0.43	9	0.27	8	0.31	16
Y	0.24	19	0.44	26	0.24	8	0.50	56	0.38	43

[a]Amino acid one-letter code.
[b]Residue propensity; cells containing values ≥ 1.01 and ≤ 1.19 and values ≥ 1.20 are shaded with light and dark grey respectively. In the latter case, numbers are boldfaces.
[c]Number of residues in the sample.

Table 3
Residue propensities in the linkers of length range 21–40.

	Actinobacteria		Alpha		Beta		Firmicutes		Gamma	
AA[a)]	P[b)]	Counts[c)]	P[b)]	Counts[c)]	P[b)]	AA[a)]	P[b)]	Counts[c)]	P[b)]	Counts[c)]
A	**1.25**	2020	1.08	1107	**1.23**	1593	0.89	1043	1.12	2186
C	0.16	18	0.10	8	0.28	32	0.37	65	0.54	123
D	0.88	638	1.17	574	1.04	585	1.05	948	0.99	1149
E	0.77	496	1.19	545	1.05	566	**1.54**	1740	1.09	1314
F	0.41	138	0.48	151	0.55	202	0.59	430	0.62	524
G	0.92	1004	0.75	520	0.72	606	0.51	546	0.60	862
H	0.84	227	0.91	154	1.02	243	**1.33**	393	1.06	501
I	0.35	157	0.59	254	0.52	247	0.81	1001	0.63	782
K	0.44	112	0.60	166	0.48	168	1.19	1377	0.69	682
L	0.65	778	0.92	776	0.90	979	0.78	1237	0.92	2051
M	0.33	73	0.65	126	0.75	179	0.56	238	0.75	357
N	0.55	139	0.55	138	0.52	170	0.93	780	0.71	641
P	**2.77**	2071	**2.68**	1163	**2.39**	1295	**2.38**	1339	**2.67**	2415
Q	0.89	322	**1.25**	356	1.11	474	**1.61**	981	**1.36**	1298
R	**1.53**	1447	1.17	691	**1.33**	997	1.15	810	**1.22**	1396
S	1.06	763	**1.20**	610	**1.34**	870	1.18	1267	**1.22**	1754
T	1.08	818	0.82	378	0.88	498	0.91	845	0.97	1119
V	0.71	717	0.87	509	0.84	619	0.72	783	1.04	1421
W	0.60	114	0.92	106	0.82	127	0.67	119	0.58	167
Y	0.37	89	0.36	70	0.31	79	0.87	585	0.45	287

[a]Amino acid one-letter code.
[b]Residue propensity; cells containing values ≥ 1.01 and ≤ 1.19 and values ≥ 1.20 are shaded with light and dark grey respectively. In the latter case, numbers are boldfaces.
[c]Number of residues in the sample.

2. Experimental design, materials and methods

Data was created from the analysis of MocR sequences taken from the most populated phyla Actinobacteria, Firmicutes, Alpha-, Beta- and Gammaproteobacteria. Sequences of the MocR regulators in each phylum were retrieved from the UniProt data bank [2] accessed on October, 2015 with the application of RPSBLAST of the BLAST suite [3] and the CDD data bank [4]. The protein sequences containing both the wHTH and AAT domains identified by RPSBLAST were considered genuine MocR regulators. Before further processing, retrieved sequences were filtered at 75% sequence identity with the program CD-HIT [5]. Multiple sequence alignments were calculated with the programs ClustalO [6] and processed with the software Jalview [7]. Linker sequences were manually extracted from the multiple sequence alignments according with the wHTH and AAT domain boundaries assigned by RPSBLAST. List of the MocR regulators possessing linkers longer than 60 residues is reported in Table 1. Residue frequency and propensities were calculated as described in [1] and are displayed in Tables 2–5 organized according to linker length and phylum class. Propensities for the entire linker set are reported in [1]. Dipeptide frequency and propensity

Table 4
Residue propensities in the linkers of length range 41–60.

	Actinobacteria		Alpha		Beta		Firmicutes		Gamma	
AA[a)]	P[b)]	Counts[c)]	P[b)]	Counts[c)]	P[b)]	AA[a)]	P[b)]	Counts[c)]	P[b)]	Counts[c)]
A	**1.31**	709	0.99	937	**1.29**	1636	0.83	775	1.10	871
C	0.22	8	0.30	22	0.23	25	0.37	52	0.53	49
D	1.03	250	1.03	463	0.94	516	1.01	732	0.90	424
E	0.85	184	0.95	401	0.93	488	1.10	997	0.89	437
F	0.62	71	0.64	183	0.45	163	0.61	356	0.47	160
G	0.94	346	0.74	475	0.86	708	0.73	623	0.68	400
H	0.71	64	0.95	149	1.12	260	**1.49**	354	**1.38**	265
I	0.51	76	0.59	236	0.67	313	0.81	801	0.64	322
K	0.43	37	1.12	287	0.58	201	0.91	841	0.89	358
L	0.64	257	0.73	572	0.76	805	0.74	936	0.83	752
M	0.61	45	0.53	95	0.75	176	0.81	278	0.57	111
N	0.63	54	0.83	191	0.57	181	**1.54**	1031	0.88	322
P	**2.20**	550	**2.41**	965	**2.30**	1219	**1.86**	839	**2.34**	863
Q	0.97	118	**1.22**	322	**1.20**	502	**1.53**	748	**1.44**	559
R	1.08	343	**1.54**	843	**1.37**	1005	1.17	662	**1.41**	660
S	**1.37**	331	**1.25**	588	**1.28**	815	**1.26**	1086	**1.39**	812
T	1.10	280	0.91	388	0.84	464	0.92	690	1.04	490
V	0.75	252	0.63	338	0.87	628	0.61	526	0.69	385
W	1.08	69	**1.51**	161	0.64	96	**3.28**	466	**1.35**	159
Y	0.32	26	0.78	138	0.33	81	0.85	458	0.61	159

[a]Amino acid one-letter code.
[b]Residue propensity; cells containing values ≥ 1.01 and ≤ 1.19 and values ≥ 1.20 are shaded with light and dark grey respectively. In the latter case, numbers are boldfaces.
[c]Number of residues in the sample.

calculations relied on the software 'compseq' of the EMBOSS suite [8]. Table 6 reports the average number of residue dyads in each group. The highest the number, the highest the reliability of the dyad propensities reported in Figs. 1–5. Average content of predicted secondary structures (obtained with the program PREDATOR [9]) are displayed in Table 7. Physicochemical properties were assigned to the amino acid residues according to the indices provided by the AAindex data bank [10] incorporated in the Interpol package [11] of the R-project library [12]. Distribution of the properties are reported as box-plots in Figs. 6–10 limited to the phyla Alphaproteobacteria, Betaproteobacteria and Gammaproteobacteria and in Figs. 11 and 12 for all the phyla considered. Box-plots for Actinobacteria and Firmicutes missing in Figs. 6–10 are to be found in [1].

The linker length distribution were analyzed within two specific MocR subfamilies: GabR [13] and PdxR [14] involved in the regulation of the synthesis of acid γ-amino butyric and pyridoxal 5′-phosphate, respectively. Sequences assigned to each of the two subgroups were retrieved from the RegPrecise data bank [15] and aligned separately (Table 8); a HMM profile [16] was calculated for each one of the multiple alignment. The profile was utilized to search for other putative GabR or

Table 5
Residue propensities in the linkers of length range 61–200.

	Actinobacteria		Alpha		Beta		Firmicutes		Gamma	
AA[a)]	P[b)]	Counts[c)]	P[b)]	Counts[c)]	P[b)]	AA[a)]	P[b)]	Counts[c)]	P[b)]	Counts[c)]
A	**1.23**	333	0.57	20	**1.67**	258	**2.06**	261	0.72	41
C	0.54	10	0.37	1	0.22	3	0.32	6	0.45	3
D	0.90	110	**1.25**	21	0.88	59	1.09	106	0.44	15
E	0.67	73	0.51	8	0.69	44	1.11	136	0.82	29
F	0.77	44	0.19	2	0.27	12	0.33	26	0.20	5
G	**1.44**	267	0.54	13	**1.54**	154	**2.08**	239	0.87	37
H	0.77	35	0.69	4	0.74	21	1.03	33	**1.88**	26
I	0.76	57	0.34	5	0.72	41	0.31	42	0.64	23
K	0.51	22	0.84	8	0.48	20	0.38	48	**1.49**	43
L	0.40	81	0.34	10	0.47	61	0.55	94	**1.46**	95
M	0.27	10	0.60	4	0.49	14	0.48	22	0.36	5
N	1.03	44	0.82	7	0.83	32	0.62	56	0.57	15
P	**1.93**	243	**1.61**	24	**1.55**	100	**2.17**	132	**2.53**	67
Q	0.77	47	**1.23**	12	1.04	53	1.04	69	0.61	17
R	1.13	180	0.54	11	**1.35**	121	**1.62**	124	0.80	27
S	**1.74**	211	0.97	17	**1.54**	119	**1.45**	169	**2.11**	89
T	**1.23**	157	0.82	13	0.77	52	0.80	81	**1.56**	53
V	0.52	89	0.85	17	0.74	65	0.71	83	0.55	22
W	0.50	16	**1.26**	5	0.71	13	**1.25**	24	0.24	2
Y	0.44	18	0.61	4	0.40	12	0.49	36	0.11	2

[a]Amino acid one-letter code.
[b]Residue propensity; cells containing values ≥ 1.01 and ≤ 1.19 and values ≥ 1.20 are shaded with light and dark grey respectively. In the latter case, numbers are boldfaces.
[c]Number of residues in the sample.

Table 6
Average number of residue pairs in each data set.

	Length intervals				
	All	**0–20**	**21–40**	**41–60**	**61–200**
Actinobacteria	53.5 ± 93.1	9.2 ± 17.6	29.2 ± 53.8	10.0 ± 16.4	5.0 ± 8.5
Alphaproteobacteria	45.7 ± 56.7	6.0 ± 9.0	20.3 ± 28.5	18.9 ± 22.4	0.5 ± 0.8
Betaproteobacteria	57.1 ± 78.2	3.2 ± 5.1	25.5 ± 35.1	25.1 ± 34.8	3.0 ± 5.8
Firmicutes	83.0 ± 63.5	6.4 ± 6.8	39.9 ± 34.8	32.4 ± 25.0	4.4 ± 6.4
Gammaproteobacteria	82.0 ± 81.9	8.7 ± 9.4	50.8 ± 54.1	20.9 ± 20.6	1.5 ± 3.5

PdxR sequences in the reference proteomes data bank available at the Hmmer web server [17]. Sequences showing an E-value smaller than 10^{-120}, were retrieved and multiply aligned. Linker sequences were extracted as described above. Length distribution were plotted and compared for the GabR and PdxR sets (Fig. 13).

Perl and R-scripts were written for data analysis, processing and display.

A

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	**1.3**	0.4	**1.1**	1.0	0.5	**1.4**	0.7	0.4	0.3	0.9	0.4	0.7	**4.2**	1.0	**1.9**	**1.5**	**1.4**	0.9	0.7	0.1
C	0.3	0.0	0.3	0.2	0.0	0.4	0.0	0.4	0.3	0.1	0.0	0.3	0.1	0.0	0.4	0.3	0.2	0.1	0.3	0.0
D	**1.2**	0.1	0.7	0.6	0.5	1.0	0.5	0.7	0.1	0.7	0.9	0.3	**1.1**	0.8	**1.2**	0.8	**1.5**	1.1	**1.3**	**1.4**
E	0.8	0.1	0.5	0.4	0.4	**1.3**	0.5	0.4	0.6	0.5	0.5	0.7	**3.1**	0.5	1.0	0.7	0.8	0.7	0.2	0.3
F	0.6	0.2	0.8	0.5	0.4	0.4	0.2	0.2	0.3	0.3	0.1	0.3	0.4	0.6	0.7	0.9	0.9	0.2	0.3	0.2
G	**1.5**	0.4	0.9	0.9	0.5	1.1	0.9	0.5	0.3	0.6	0.4	0.5	**2.2**	0.5	**1.4**	**1.1**	**1.1**	1.0	1.0	0.4
H	0.9	0.2	**1.2**	0.5	0.2	0.7	0.3	0.8	0.1	0.7	0.4	0.4	**1.6**	0.7	0.9	**1.1**	1.1	0.4	0.0	0.2
I	0.6	0.1	0.7	0.6	0.6	0.4	0.3	0.2	0.0	0.2	0.0	0.5	**1.1**	0.5	0.7	0.3	0.2	0.1	0.0	0.0
K	0.4	0.7	0.2	0.3	0.2	0.2	0.4	0.4	0.3	0.3	0.2	0.6	**1.9**	0.2	0.5	0.2	0.3	0.4	0.2	0.1
L	0.5	0.1	0.5	0.5	0.7	0.6	0.9	0.3	0.3	0.3	0.4	0.3	**2.0**	0.3	0.8	1.0	0.5	0.3	0.5	0.2
M	0.3	0.0	0.4	0.3	0.4	0.2	0.3	0.1	0.0	0.4	0.3	0.2	0.9	0.5	0.6	0.4	0.2	0.2	0.2	0.2
N	0.7	0.0	**0.8**	0.4	0.3	0.4	0.2	0.3	0.4	0.5	1.0	0.5	0.9	0.3	1.0	0.6	1.0	0.4	0.6	0.5
P	**3.8**	0.8	**2.1**	**2.0**	0.6	**2.5**	**2.5**	**1.1**	**1.6**	**1.5**	**1.3**	**1.2**	**5.9**	**2.3**	**4.9**	**2.7**	**2.5**	**2.0**	**1.3**	0.4
Q	1.1	0.0	0.7	**1.4**	0.1	0.9	0.6	0.5	0.4	0.5	0.5	0.6	**2.0**	0.6	**1.2**	0.9	**1.9**	0.5	0.2	0.2
R	**1.7**	0.1	1.1	0.9	0.5	**1.3**	**0.8**	0.8	0.6	0.7	0.5	0.6	**4.1**	1.1	**1.4**	**1.7**	**2.0**	1.0	0.5	0.8
S	**1.5**	0.2	1.0	0.8	0.7	**1.3**	**1.2**	0.5	0.6	0.7	0.7	**1.1**	**2.3**	0.8	**1.7**	**1.4**	**1.4**	0.9	**2.1**	0.4
T	**1.5**	0.2	0.8	0.9	0.4	0.9	0.8	0.6	0.6	0.7	0.4	0.6	**2.9**	0.9	**1.5**	**1.4**	**1.4**	0.8	0.5	0.5
V	0.8	0.2	0.8	0.5	0.6	0.5	0.5	0.2	0.2	0.3	0.3	0.2	**2.4**	0.5	1.0	0.8	0.6	0.4	0.3	0.0
W	0.5	0.2	0.3	0.6	0.1	0.2	0.5	0.2	0.3	0.5	0.5	0.2	0.3	**6.1**	0.6	**1.3**	0.6	0.0	0.0	0.3
Y	0.5	0.0	0.1	0.3	0.0	0.5	**1.3**	0.0	0.1	0.1	0.0	0.4	0.6	0.3	0.5	1.1	0.1	0.2	0.0	0.0

B

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	1.0	0.3	**1.1**	**1.1**	0.5	0.9	0.9	0.8	0.8	0.8	0.9	0.6	**3.6**	**1.1**	**1.4**	**1.3**	0.9	1.0	0.6	0.6
C	0.1	0.4	0.2	0.2	0.1	0.2	0.0	0.1	0.2	0.2	0.4	0.2	0.1	0.8	0.2	0.4	0.3	0.1	0.0	0.2
D	1.0	0.1	**1.1**	0.7	0.8	0.5	0.6	0.6	0.8	0.8	0.8	0.9	**2.7**	**1.2**	**1.8**	**1.4**	**1.3**	1.0	**5.2**	0.8
E	**1.3**	0.3	1.0	0.8	0.5	0.8	1.0	0.8	0.7	0.6	1.1	**1.1**	**3.6**	**1.2**	**1.6**	1.1	**1.3**	1.1	0.4	0.3
F	0.5	0.0	0.6	0.4	0.1	0.5	0.4	0.3	0.6	0.2	0.4	0.2	**1.4**	0.4	0.9	0.7	0.6	0.3	**2.9**	0.1
G	0.8	0.2	0.8	**1.3**	0.3	0.9	0.8	0.4	0.7	0.6	0.6	0.5	**2.6**	**1.2**	1.0	0.7	0.9	0.7	0.2	0.2
H	1.0	0.0	0.8	0.7	0.4	0.2	0.9	0.6	0.8	1.0	**1.4**	0.9	**1.5**	0.9	**1.4**	**1.2**	0.9	0.9	0.8	**1.6**
I	0.5	0.1	0.9	0.7	0.2	0.7	0.5	0.2	0.4	0.3	0.5	0.4	**1.3**	0.8	0.6	0.8	0.4	0.4	0.3	0.2
K	0.7	0.0	0.7	0.8	0.5	0.5	0.9	0.3	0.3	1.0	0.5	0.6	**2.7**	0.9	0.9	0.8	0.8	0.4	0.4	0.4
L	0.9	0.2	**1.2**	1.0	0.3	0.5	0.8	0.3	0.6	0.5	0.5	0.5	**2.3**	0.9	0.8	**1.2**	0.6	0.4	**1.7**	0.3
M	0.5	0.0	0.9	0.5	0.4	0.6	0.4	0.4	0.4	0.6	0.5	0.2	1.1	0.6	0.6	0.5	0.5	0.3	0.3	0.4
N	0.8	0.4	0.6	0.5	0.4	0.4	0.7	0.9	0.5	**0.8**	0.6	0.1	1.1	0.7	0.7	0.7	0.7	0.4	**3.1**	0.5
P	**2.6**	0.4	**2.3**	**2.9**	**1.8**	**1.7**	**2.7**	**1.5**	**2.2**	**2.7**	**1.5**	**1.3**	**6.0**	**3.0**	**3.7**	**2.8**	**1.9**	**1.9**	1.1	**1.5**
Q	**1.3**	0.2	**1.3**	0.9	0.4	0.8	**1.3**	0.8	1.1	**1.2**	0.6	1.1	**2.6**	0.9	**1.8**	**2.4**	**1.4**	0.8	0.4	0.5
R	**1.4**	0.3	1.1	**1.5**	**1.3**	**1.3**	**1.2**	0.6	1.0	**1.2**	0.9	1.0	**3.5**	**2.6**	**1.5**	**1.5**	**1.3**	0.8	0.5	1.0
S	**1.5**	0.5	1.1	**1.4**	0.7	0.9	**1.2**	0.5	**1.5**	**1.5**	**1.5**	0.8	**2.0**	**1.2**	**2.3**	**1.2**	0.8	0.9	0.9	**1.2**
T	0.9	0.0	0.9	**1.4**	0.4	0.8	0.8	0.5	0.8	0.6	0.3	0.7	**1.8**	0.9	**1.6**	**1.1**	0.7	0.6	0.2	0.6
V	0.8	0.0	**1.7**	0.7	0.2	0.5	0.7	0.3	0.6	0.3	0.3	0.5	**1.8**	0.9	0.8	**1.3**	0.5	0.3	0.7	0.1
W	**1.2**	0.0	0.6	0.4	0.3	0.8	0.8	**1.4**	0.5	0.5	**1.2**	1.0	0.9	**1.7**	**1.6**	**1.7**	1.0	**3.1**	0.2	0.4
Y	0.7	0.2	0.3	0.7	0.3	0.6	0.5	0.4	**1.1**	0.4	0.2	0.1	**1.5**	0.8	1.0	0.4	0.4	0.2	0.2	0.1

C

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	**1.6**	0.3	1.1	1.1	0.5	**1.1**	**1.2**	0.8	0.7	0.9	**1.2**	0.7	**3.6**	**1.5**	**1.9**	**1.7**	**1.2**	0.9	0.7	0.4
C	0.2	0.2	0.5	0.3	0.0	0.2	0.0	0.1	0.0	0.2	0.2	0.3	0.6	0.8	0.2	0.2	0.3	0.3	0.0	0.2
D	**1.1**	0.0	0.7	0.9	0.8	0.8	**1.3**	1.0	0.4	0.8	1.0	0.6	**1.7**	1.1	**1.2**	**1.3**	**1.4**	**1.5**	**1.6**	0.2
E	1.0	0.1	0.8	0.7	0.5	0.9	1.0	0.7	0.5	0.7	0.9	0.5	**3.3**	**1.3**	**1.3**	**1.4**	**1.2**	0.9	0.7	0.3
F	0.7	0.3	0.6	0.7	0.1	0.3	**1.2**	0.3	0.3	0.3	0.5	0.3	0.6	0.5	0.9	0.6	0.3	0.4	0.6	0.1
G	**1.3**	0.1	0.7	1.0	0.3	0.7	1.1	0.7	0.4	0.9	1.0	0.4	**2.0**	1.0	**1.1**	**1.2**	0.9	0.9	0.3	0.1
H	**1.8**	0.6	**1.3**	**1.4**	0.4	1.0	0.9	0.6	0.8	0.7	0.5	0.7	**1.6**	**1.2**	**1.1**	**1.2**	**1.2**	0.9	0.3	0.3
I	0.7	0.0	0.9	0.4	0.5	0.4	**1.2**	0.2	0.3	0.5	0.2	0.5	0.9	0.4	0.9	1.0	0.3	0.6	0.6	0.1
K	0.5	0.6	0.4	0.4	0.1	0.3	1.0	0.2	0.2	0.3	0.3	0.2	**1.7**	0.6	0.9	0.7	0.6	0.6	0.1	0.3
L	0.9	0.1	0.7	0.9	0.3	0.6	0.7	0.4	0.5	0.5	0.7	0.5	**2.4**	0.9	0.9	**1.4**	0.6	0.6	1.0	0.3
M	0.6	0.2	0.7	0.6	**2.2**	0.5	0.7	0.1	0.2	0.6	0.5	0.2	**2.0**	0.8	1.0	0.8	0.8	0.3	0.4	0.2
N	0.8	0.9	0.6	0.6	0.4	0.3	0.5	0.9	0.1	0.6	**1.4**	0.4	0.5	0.3	0.6	0.6	0.6	0.4	**1.1**	0.6
P	**3.3**	0.5	**2.6**	**2.3**	0.6	**2.2**	**1.6**	**1.2**	**1.3**	**2.0**	**1.7**	0.9	**4.4**	**2.0**	**3.6**	**2.5**	**1.7**	**1.8**	0.6	0.7
Q	**1.3**	0.1	0.9	0.8	1.1	0.8	**1.8**	0.7	0.5	**1.1**	0.9	0.6	**2.5**	1.6	**1.7**	1.2	0.8	1.0	**1.8**	1.0
R	**1.6**	0.4	0.9	**1.2**	0.5	**1.9**	**1.2**	0.7	0.8	**1.1**	1.1	**1.1**	**3.2**	**1.6**	**1.5**	**1.8**	**1.3**	**1.3**	0.7	0.4
S	**1.7**	0.4	**1.3**	**1.6**	0.4	**1.2**	**1.3**	1.0	1.0	**1.2**	**1.5**	0.8	**2.3**	**1.4**	**2.0**	**1.3**	**1.5**	0.9	0.5	0.5
T	**1.2**	0.4	0.5	1.0	0.3	0.7	0.9	0.3	0.5	0.8	0.8	0.4	**1.8**	**1.4**	**1.4**	1.0	0.7	0.6	0.5	0.1
V	1.0	0.1	**1.4**	0.4	0.9	0.7	0.9	0.3	0.5	0.7	0.4	0.4	**1.8**	0.8	0.9	**1.5**	0.5	0.4	**1.2**	0.2
W	0.1	0.0	1.0	1.1	0.1	0.3	0.3	0.4	0.3	**1.4**	0.7	0.9	0.3	**1.3**	0.6	**1.2**	0.2	0.5	0.0	0.7
Y	0.4	0.2	0.1	0.5	0.2	0.2	0.4	0.2	0.1	0.4	0.6	0.1	0.8	0.6	0.2	0.5	0.0	0.6	0.1	0.0

D

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	**1.3**	0.5	1.0	**1.4**	0.5	1.1	0.9	0.5	0.7	0.5	0.5	0.6	**2.6**	1.0	**1.1**	**1.4**	0.8	0.7	**1.1**	0.9
C	0.3	0.5	0.6	0.9	0.2	0.1	0.7	0.3	0.3	0.5	0.3	0.3	0.3	0.6	0.5	0.2	0.3	0.1	0.8	0.1
D	0.9	1.0	0.9	1.1	0.6	0.7	**1.2**	**1.6**	**0.8**	1.0	1.0	0.9	**1.4**	1.1	1.0	**1.4**	1.0	0.9	**5.2**	**1.1**
E	**1.3**	0.3	**1.3**	**1.8**	**1.3**	**1.5**	**1.7**	**1.2**	**1.5**	1.0	0.8	**1.3**	**3.1**	**1.8**	**1.3**	**1.6**	**1.2**	0.8	**1.2**	**1.3**
F	0.5	0.2	0.7	0.8	0.2	0.2	0.8	0.4	0.7	0.5	0.3	0.6	**1.1**	**1.5**	0.9	0.7	0.6	0.3	0.4	0.3
G	0.8	0.3	0.6	0.9	0.4	0.7	0.9	0.7	0.6	0.6	1.0	0.6	**1.1**	**1.1**	0.9	**1.2**	0.6	0.7	0.6	0.4
H	**1.5**	0.1	**1.1**	**1.8**	0.9	0.8	**1.6**	**1.4**	**2.0**	1.0	**1.6**	**1.5**	**1.4**	**2.0**	**1.9**	**1.4**	**1.8**	**1.3**	**2.2**	**1.7**
I	0.5	0.3	0.7	1.1	0.3	0.2	**1.7**	0.4	0.9	0.4	0.4	**1.7**	**1.4**	**2.0**	**1.2**	0.9	0.6	0.4	1.1	0.4
K	0.8	0.5	1.0	**1.3**	0.4	0.6	**2.0**	0.8	1.2	0.7	0.8	**1.3**	**2.0**	**1.5**	**1.1**	**1.3**	**1.1**	0.7	**2.4**	0.7
L	0.6	0.2	1.0	1.0	0.4	0.3	**1.2**	0.4	0.5	0.5	0.7	0.7	**2.6**	1.0	0.7	1.0	0.6	0.3	0.5	0.6
M	0.4	0.3	0.7	0.9	0.2	0.3	1.1	0.5	0.6	0.4	0.5	0.7	**1.4**	0.7	0.6	0.7	0.5	0.4	0.7	0.3
N	0.8	0.4	**1.2**	**1.4**	0.9	0.4	**1.8**	**1.2**	**1.3**	**1.1**	**1.1**	**1.9**	**1.2**	**1.7**	**1.4**	**1.4**	**1.7**	0.8	**5.6**	1.0
P	**2.6**	0.7	**2.2**	**1.8**	**1.2**	**1.6**	**1.6**	**1.3**	**1.8**	**2.0**	**1.2**	**3.7**	**5.9**	**2.4**	**2.7**	**2.6**	**1.8**	**2.2**	**2.5**	**1.6**
Q	**1.4**	0.0	**1.7**	**1.9**	1.1	1.0	**2.2**	**1.2**	**1.9**	**1.1**	**1.2**	**1.3**	**3.5**	**1.8**	**1.8**	**1.6**	**1.9**	1.0	**1.7**	**1.4**
R	**1.2**	0.3	1.1	**1.5**	0.5	0.7	**1.5**	1.0	**1.2**	0.8	**1.2**	1.0	**2.6**	**1.3**	**1.2**	**1.8**	**1.2**	0.8	0.9	1.0
S	**1.4**	0.9	**1.4**	**1.5**	0.9	1.0	**1.8**	0.9	**1.2**	1.0	**1.2**	**1.4**	**2.7**	**1.7**	**1.6**	**1.5**	**1.2**	0.8	**2.4**	**1.2**
T	0.8	0.6	1.0	**1.3**	0.6	0.6	**1.2**	0.9	1.1	0.6	0.5	1.0	**2.0**	**1.3**	0.7	**1.3**	0.6	0.5	**3.0**	1.0
V	0.4	0.3	0.6	1.1	0.3	0.3	1.1	0.3	0.8	0.4	0.4	0.7	**1.8**	**1.7**	0.9	0.9	0.6	0.3	0.4	0.5
W	0.5	0.5	**1.3**	**1.9**	0.6	**4.5**	**3.5**	**1.5**	0.4	0.3	**4.3**	1.0	**4.1**	**2.6**	**3.4**	**1.1**	0.2	0.0	0.7	
Y	0.7	0.1	0.6	0.9	0.5	0.4	**1.2**	**0.8**	1.0	1.0	0.7	0.6	**1.3**	**1.4**	1.0	0.9	0.6	**1.4**	0.2	0.2

E

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	**1.2**	0.7	0.8	1.0	0.6	0.9	**1.3**	0.7	0.8	0.9	1.0	0.7	**3.5**	**1.6**	**1.7**	**1.6**	**1.1**	1.0	0.4	0.4
C	0.5	0.9	0.6	0.7	0.2	0.3	0.5	0.1	0.9	0.5	0.2	0.2	0.9	0.7	0.8	0.5	0.5	0.5	0.2	0.3
D	0.9	0.3	0.7	1.0	0.7	0.4	**1.2**	1.1	0.6	**1.3**	1.0	0.7	**1.3**	**1.4**	**1.2**	**1.2**	**1.4**	**1.3**	1.3	0.6
E	0.9	0.4	0.7	0.6	0.6	1.0	**1.7**	1.0	0.8	1.1	0.9	0.8	**2.6**	**1.4**	**1.3**	**1.4**	**1.3**	1.1	0.7	0.5
F	0.6	0.2	0.9	0.8	0.3	0.3	0.8	0.4	0.5	0.5	0.4	0.6	1.0	**1.2**	0.9	0.7	0.5	0.5	0.4	0.2
G	0.8	**0.8**	0.7	0.7	0.5	0.6	1.0	0.5	0.4	0.6	**1.3**	0.5	**1.6**	1.0	0.7	0.8	0.6	0.7	0.3	0.3
H	**1.7**	0.3	1.0	**1.2**	0.9	0.9	**1.4**	0.9	0.9	**1.3**	1.0	1.0	**1.9**	**1.3**	1.1	**1.7**	1.0	**1.3**	0.8	1.0
I	0.5	0.6	0.8	0.8	0.5	0.3	1.0	0.2	0.5	0.5	0.9	1.0	**1.3**	**1.1**	0.5	1.0	0.5	0.4	0.4	0.3
K	0.7	0.3	0.5	0.6	0.3	0.4	0.8	0.7	0.7	0.6	0.4	0.5	**2.3**	0.8	0.9	0.9	0.9	0.8	0.9	0.3
L	0.9	0.3	1.0	1.0	0.5	0.4	1.0	0.5	0.7	0.7	0.4	0.9	**3.1**	**1.2**	0.7	**1.3**	0.7	0.6	**1.1**	0.6
M	0.8	1.0	0.9	1.0	0.2	0.5	0.7	0.3	0.5	0.6	0.7	0.4	**1.4**	0.7	0.7	0.8	0.7	0.4	0.5	0.3
N	0.8	0.4	0.9	0.6	0.5	0.3	**1.1**	0.7	0.7	0.7	0.6	0.6	1.1	0.8	0.7	1.0	0.9	0.7	**1.8**	0.7
P	**3.3**	**1.5**	**2.5**	**2.5**	0.9	**2.0**	**2.1**	**1.2**	**2.0**	**2.2**	**1.8**	**1.3**	**5.5**	**3.2**	**3.5**	**3.0**	**2.3**	**2.9**	0.4	0.9
Q	**1.4**	0.6	**1.2**	**1.3**	1.0	0.8	**1.6**	1.0	1.0	1.0	**1.3**	1.1	**3.3**	**1.6**	**1.7**	**1.4**	**1.7**	**1.6**	**3.2**	1.0
R	**1.4**	0.4	0.9	**1.2**	0.8	**1.2**	**1.3**	0.5	0.8	1.1	0.9	0.8	**4.2**	**1.8**	**1.6**	**1.6**	**1.2**	0.9	0.6	0.5
S	**1.5**	0.7	**1.2**	**1.4**	0.5	0.9	**1.8**	0.8	**1.2**	**1.2**	1.0	0.8	**2.8**	**1.9**	**2.0**	**1.5**	**1.7**	**1.2**	0.5	0.4
T	**1.2**	0.6	0.8	**1.1**	0.5	0.7	**1.1**	0.6	0.7	0.9	0.5	0.7	**2.2**	**1.5**	**1.5**	**1.3**	1.1	0.8	0.6	0.3
V	1.0	0.4	**1.2**	0.9	1.0	0.5	0.9	0.6	0.7	1.0	0.8	0.8	**2.0**	**1.4**	**1.1**	**1.5**	0.9	0.5	0.4	0.5
W	0.6	0.3	**1.6**	**1.3**	0.3	0.6	0.5	0.9	0.4	0.4	**1.2**	0.5	0.5	1.1	1.1	1.1	0.9	0.4	0.7	0.0
Y	0.6	0.2	0.5	0.5	0.1	0.3	0.8	0.2	0.3	0.6	0.4	0.8	0.9	1.0	0.5	0.6	0.3	0.4	0.2	0.2

Fig. 1. Dipeptide propensity for the entire set of linkers. Vertical and horizontal sides of each matrix indicate the N- and C-side residue of each dyad, respectively. Cells containing propensity values ≥ 1.1 and ≤ 1.99 or ≥ 2.0 and ≤ 3.99 or ≥ 4.0 are shaded with very light, light or dark grey respectively and numbers therein contained are boldfaced. A, B, C, D and E denote propensities for Actinobacteria, Alphaproteobacteria, Betaproteobacteria, Firmicutes and Gammaproteobacteria, respectively.

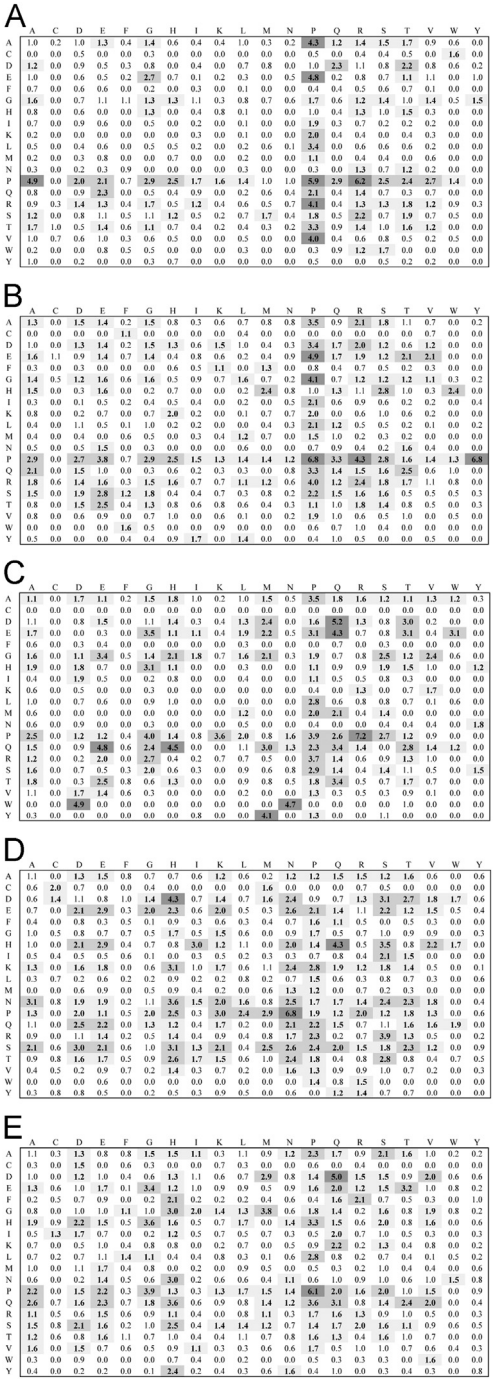

Fig. 2. Dipeptide propensity for the 0–20 residue length linker set. Interpretation of figure refers to legend to Fig. 1.

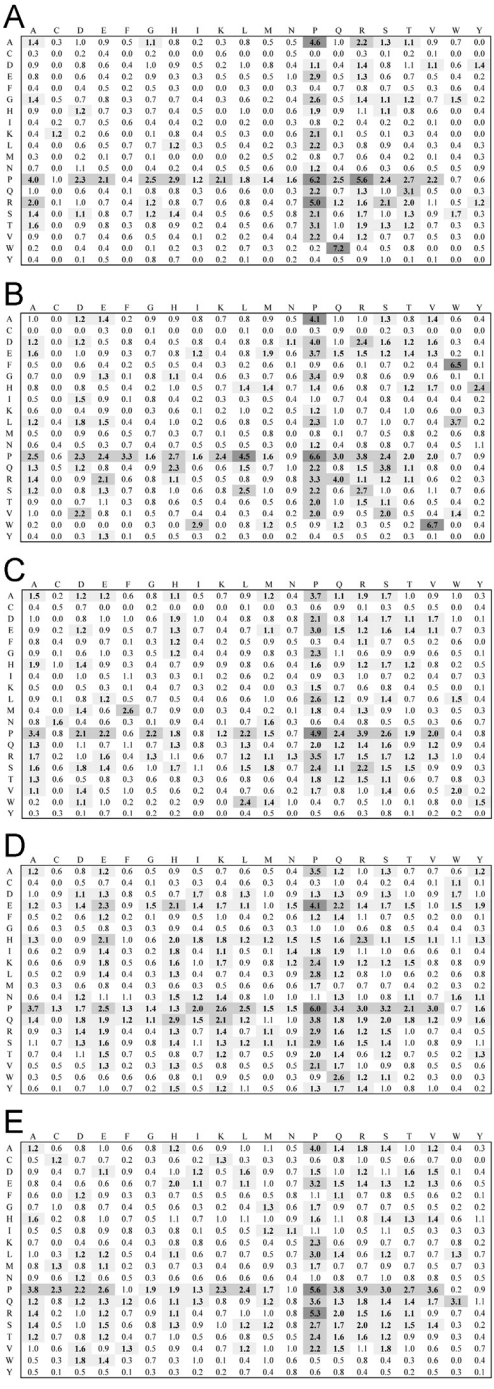

Fig. 3. Dipeptide propensity for the 21–40 residue length linker set. Interpretation of figure refers to legend to Fig. 1.

A

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	1.5	0.2	1.2	1.0	0.7	1.6	0.6	0.6	0.4	1.1	0.3	1.2	3.8	0.4	1.7	1.4	1.9	0.9	0.5	0.1
C	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.8	0.0	0.4	0.0	0.0
D	2.1	0.6	0.6	0.5	0.3	0.8	0.4	0.8	0.0	0.3	1.4	0.2	1.0	0.1	0.9	0.9	2.6	1.1	4.3	1.9
E	0.9	0.0	0.5	0.3	1.5	0.7	0.7	0.9	0.9	0.6	1.0	0.4	2.7	1.1	0.7	0.9	0.7	1.1	0.0	0.2
F	0.8	0.0	1.8	0.7	1.6	0.3	0.0	0.0	1.4	0.2	0.7	0.0	0.6	0.4	0.7	0.4	1.4	0.1	0.0	0.0
G	1.3	0.6	1.5	0.6	0.6	0.8	1.3	0.5	0.2	0.7	1.0	0.3	1.8	0.3	1.2	0.9	0.8	1.4	0.4	0.2
H	1.2	1.1	0.9	0.2	0.0	0.7	0.7	1.4	0.0	0.4	1.9	0.6	1.5	0.7	0.5	0.7	1.2	0.0	0.0	0.0
I	0.6	0.0	0.8	0.9	1.1	0.1	0.4	0.0	0.0	0.3	0.0	1.6	0.7	1.7	0.6	0.4	0.4	0.1	0.0	0.0
K	0.6	0.0	0.0	0.0	0.9	0.3	0.0	0.3	0.0	0.3	0.0	1.1	1.3	0.3	0.7	0.4	0.0	0.8	0.0	0.0
L	0.9	0.0	0.4	0.5	1.0	0.3	0.8	0.1	0.1	0.2	0.9	0.5	1.2	0.6	0.9	1.6	0.4	0.5	1.1	0.0
M	0.3	0.0	0.9	0.4	0.0	0.4	0.7	0.3	0.0	1.4	0.0	0.0	1.3	1.0	0.9	0.5	0.2	0.2	0.0	0.0
N	0.7	0.0	0.2	0.3	0.4	0.4	0.6	0.3	0.5	0.5	3.5	2.0	0.3	0.4	1.0	0.8	1.7	0.3	0.8	0.0
P	2.8	0.0	2.0	1.9	0.3	2.6	1.7	0.4	0.6	1.2	1.9	0.4	5.8	1.1	2.9	3.2	2.2	1.3	1.9	0.5
Q	1.2	0.0	1.0	4.2	0.0	1.1	0.3	0.8	0.0	0.5	2.2	1.1	1.8	0.7	0.8	0.8	0.8	0.3	1.0	0.0
R	1.6	0.0	1.0	0.9	0.2	0.9	0.9	0.5	0.5	0.5	1.1	0.3	2.4	1.4	1.1	1.6	1.7	0.9	0.0	0.5
S	1.8	0.9	0.8	0.6	0.6	1.8	0.4	0.7	1.0	0.8	0.5	1.4	2.7	1.4	1.4	2.4	0.6	1.2	5.8	1.0
T	1.3	0.0	1.1	0.9	0.4	0.9	0.4	1.2	1.3	1.1	0.0	0.2	2.3	0.8	1.1	1.9	1.3	0.5	1.5	0.6
V	0.7	0.3	1.3	0.5	1.0	0.3	0.4	0.4	0.3	0.6	0.0	0.2	2.2	0.7	0.8	1.0	0.6	0.4	0.2	0.0
W	1.4	0.0	0.3	1.4	0.0	0.0	2.0	0.0	0.7	0.1	0.9	0.6	0.6	11.1	0.4	2.5	0.7	0.0	0.0	0.0
Y	0.3	0.0	0.0	0.2	0.0	0.1	0.6	0.0	0.0	0.4	0.0	0.5	1.6	0.0	0.0	0.7	0.2	0.8	0.0	0.0

B

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0.8	0.7	1.0	0.7	0.7	0.7	0.9	1.0	0.9	0.8	0.9	0.8	3.1	1.3	1.5	1.2	1.0	0.6	0.8	0.9
C	0.1	0.8	0.5	0.3	0.0	0.4	0.0	0.3	0.5	0.1	0.9	0.5	0.0	0.9	0.5	0.5	0.3	0.0	0.0	0.6
D	0.9	0.3	0.9	0.7	1.0	0.3	0.5	0.4	0.9	0.7	0.7	1.0	1.1	1.2	1.2	1.2	1.6	0.4	12.2	1.5
E	0.9	0.3	0.9	0.5	0.7	0.8	1.4	0.5	1.1	0.5	0.5	1.7	3.2	0.8	1.6	0.8	1.0	0.5	0.8	0.5
F	0.7	0.0	0.7	0.6	0.1	0.6	0.4	0.2	0.8	0.3	0.0	0.4	2.1	0.2	1.8	0.8	1.2	0.3	0.0	0.1
G	0.7	0.3	0.6	1.1	0.3	0.8	0.6	0.3	0.8	0.7	0.5	0.4	1.5	1.7	1.3	0.6	0.6	0.6	0.4	0.2
H	1.1	0.0	1.1	0.6	0.5	0.2	0.9	1.0	1.2	0.8	1.1	1.3	1.9	0.8	2.1	1.4	0.6	0.3	1.1	1.2
I	0.5	0.2	0.6	0.6	0.3	0.8	0.7	0.1	0.6	0.4	0.7	0.3	1.5	0.8	0.8	0.8	0.5	0.6	0.2	0.1
K	0.9	0.0	1.1	0.7	1.3	0.6	1.0	0.6	0.4	1.1	0.7	0.6	4.1	1.3	1.4	1.1	0.7	0.4	1.0	0.5
L	0.7	0.1	0.6	0.6	0.3	0.5	0.7	0.4	0.8	0.3	0.6	0.6	2.5	0.7	0.9	1.6	0.6	0.4	0.1	0.4
M	0.5	0.0	1.0	0.6	0.2	0.4	0.3	0.1	0.7	0.4	0.0	0.4	1.3	0.9	0.7	0.7	0.4	0.4	0.0	0.0
N	1.1	0.5	0.7	0.4	0.2	0.3	0.9	1.7	0.8	1.0	1.1	0.1	1.0	1.0	0.8	0.8	0.4	0.5	6.3	0.2
P	2.7	0.3	2.1	3.1	0.8	1.5	2.8	1.5	2.1	1.3	1.4	1.8	5.4	3.0	3.6	3.3	1.9	1.9	1.4	0.6
Q	1.2	0.0	1.4	0.9	0.6	0.8	0.6	1.1	1.7	1.1	0.7	1.1	2.9	0.8	2.3	1.2	1.4	0.8	0.6	0.9
R	1.4	0.4	1.2	0.8	2.4	1.9	1.3	0.7	1.7	1.6	0.9	1.3	3.4	1.6	1.7	1.8	1.3	1.1	0.6	2.2
S	1.8	0.9	1.2	1.0	0.6	0.9	1.6	0.5	2.6	0.9	0.5	0.9	1.8	1.8	2.2	1.3	1.0	0.9	0.8	2.1
T	1.1	0.0	1.0	1.4	0.5	0.7	0.8	0.6	1.0	0.7	0.1	0.9	1.8	0.7	1.6	1.0	0.8	0.7	0.2	1.1
V	0.5	0.0	1.4	0.6	0.3	0.4	0.6	0.4	0.8	0.5	0.2	0.9	1.5	0.7	1.1	0.7	0.5	0.4	0.2	0.1
W	2.6	0.0	1.2	0.9	0.3	1.3	1.9	0.2	0.9	0.4	1.7	1.7	1.1	2.7	3.1	3.3	2.1	0.2	0.6	0.4
Y	1.1	0.5	0.4	0.3	0.5	0.6	0.3	0.1	2.4	0.3	0.4	0.2	3.1	1.1	1.8	0.7	0.7	0.3	0.4	0.2

C

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	1.6	0.3	0.9	0.9	0.4	1.2	1.2	0.9	0.7	0.8	1.3	0.9	3.5	1.8	1.9	1.7	1.4	0.9	0.4	0.4
C	0.2	0.0	0.4	0.6	0.0	0.1	0.0	0.2	0.0	0.3	0.5	0.4	0.8	0.9	0.2	0.1	0.2	0.0	0.0	0.0
D	1.3	0.0	0.5	0.7	0.6	0.9	0.9	1.0	0.3	0.7	1.0	0.6	1.4	0.7	0.9	0.8	1.3	1.4	2.3	0.3
E	1.0	0.0	0.5	0.6	0.6	0.6	0.8	0.7	0.6	0.7	0.5	0.3	3.8	0.7	1.6	1.1	0.8	0.9	0.3	0.3
F	0.6	0.2	0.4	0.7	0.1	0.3	1.3	0.4	0.4	0.1	0.2	0.1	1.0	0.6	0.8	0.4	0.3	0.6	0.8	0.1
G	1.4	0.2	0.6	0.8	0.1	0.7	0.9	0.8	0.5	0.7	1.2	0.5	1.8	1.0	1.5	1.0	0.7	0.9	0.1	0.2
H	1.8	0.3	1.3	2.0	0.6	1.1	1.0	0.5	0.7	0.7	0.3	0.9	1.7	1.6	1.3	0.8	1.2	1.0	0.5	0.2
I	1.0	0.0	0.8	0.4	0.1	0.3	2.2	0.3	0.4	0.5	0.2	0.8	1.0	0.5	0.9	1.5	0.4	0.6	0.6	0.0
K	0.5	1.2	0.2	0.4	0.1	0.2	1.2	0.1	0.2	0.3	0.5	0.1	2.2	0.7	1.1	0.6	0.7	0.6	0.2	0.6
L	0.9	0.1	0.6	0.8	0.1	0.4	0.8	0.4	0.5	0.4	0.5	0.5	2.3	0.8	0.9	1.5	0.5	0.8	0.6	0.2
M	0.8	0.5	0.3	0.7	2.1	0.4	0.4	0.3	0.2	0.7	1.0	0.1	2.2	1.2	0.8	0.8	0.7	0.4	0.5	0.3
N	0.8	0.3	0.8	0.6	0.6	0.4	0.2	1.3	0.2	0.5	1.6	0.3	0.4	0.2	0.5	0.5	0.6	0.6	1.3	0.2
P	3.4	0.2	3.3	2.6	0.7	2.0	1.7	1.5	1.3	1.9	2.0	1.0	4.3	1.7	3.1	2.4	1.5	1.9	0.8	0.6
Q	1.2	0.3	0.7	0.5	1.2	0.7	1.9	0.7	0.7	1.0	1.2	0.4	3.1	1.9	1.9	1.0	0.5	0.8	2.9	1.6
R	1.6	0.6	1.0	0.8	0.8	2.2	1.5	0.9	0.8	1.1	1.1	1.0	2.9	1.5	1.6	1.8	1.3	1.4	0.6	0.5
S	1.6	0.3	1.1	2.1	0.4	1.0	1.0	1.0	1.4	1.0	1.4	0.9	2.1	1.8	2.1	1.2	1.5	1.0	0.1	0.5
T	1.0	0.2	0.5	0.9	0.2	0.7	1.1	0.4	0.5	0.9	1.1	0.3	1.9	1.5	1.4	0.9	0.8	0.6	0.1	0.0
V	1.0	0.1	1.4	0.2	0.9	0.9	1.1	0.3	0.6	0.8	0.4	0.5	2.0	0.9	0.9	1.8	0.4	0.4	0.5	0.3
W	0.1	0.0	0.5	1.2	0.0	0.5	0.5	0.1	0.2	0.7	0.0	0.4	0.3	1.9	0.7	1.6	0.1	0.1	0.0	0.0
Y	0.4	0.0	0.1	0.3	0.2	0.4	0.4	0.2	0.1	0.4	0.0	0.0	0.9	0.4	0.2	0.1	0.0	1.1	0.0	0.0

D

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	1.1	0.5	0.9	1.3	0.3	1.1	0.8	0.5	0.7	0.5	0.5	0.9	1.7	0.6	1.0	1.2	0.7	0.6	1.4	0.5
C	0.1	0.8	0.7	1.4	0.1	0.2	1.0	0.3	0.1	0.5	0.0	0.2	0.4	0.2	0.7	0.1	0.3	0.0	0.0	0.2
D	0.7	1.0	0.6	0.9	0.5	0.6	1.2	1.8	0.7	0.8	1.0	0.8	1.7	0.8	0.7	1.2	0.8	0.7	11.0	1.3
E	1.4	0.3	1.0	1.0	2.0	1.1	1.2	1.2	1.3	1.1	0.6	0.9	1.9	1.4	1.1	1.3	1.0	0.4	0.7	0.9
F	0.5	0.3	0.8	0.6	0.2	0.3	0.7	0.3	0.5	0.7	0.4	0.6	0.9	1.7	0.8	0.8	0.7	0.3	0.9	0.6
G	0.7	0.1	0.5	0.8	0.3	0.8	0.5	1.0	0.5	0.5	1.8	0.8	0.8	1.1	1.1	1.3	0.6	0.7	1.1	0.3
H	1.6	0.3	1.3	1.4	1.1	1.2	1.5	0.8	2.6	1.0	2.7	1.7	1.3	2.0	1.9	1.4	2.2	1.2	3.7	2.4
I	0.3	0.5	0.5	0.9	0.3	0.3	2.2	0.4	0.3	0.7	2.6	1.2	2.6	1.5	0.6	0.6	0.2	2.6	0.6	0.6
K	0.8	0.4	1.0	0.7	0.3	0.8	2.4	0.8	0.7	0.6	0.9	1.3	1.6	1.2	1.1	1.4	0.8	0.6	5.1	0.5
L	0.7	0.1	1.2	0.8	0.5	0.2	1.1	0.4	0.4	0.7	1.1	0.6	2.4	0.9	0.8	1.0	0.6	0.3	0.4	0.3
M	0.6	0.4	1.0	1.1	0.0	0.2	1.7	0.7	1.0	0.3	0.6	0.7	1.3	0.8	0.6	1.0	0.5	0.7	1.5	0.5
N	0.5	0.5	1.2	1.9	0.7	0.3	2.1	1.3	1.2	1.6	1.5	3.0	1.3	2.4	1.9	2.0	2.4	0.7	12.2	1.2
P	1.3	0.3	3.1	1.3	1.4	1.2	1.8	0.8	0.7	1.4	0.7	6.2	6.4	1.6	2.4	2.3	1.5	1.4	4.8	1.9
Q	1.5	0.0	1.5	2.0	1.3	0.7	1.7	1.0	1.9	1.3	1.7	1.5	3.4	1.9	1.8	1.4	2.3	0.6	2.8	1.7
R	1.3	0.5	0.7	1.1	0.5	1.0	1.8	1.6	1.1	0.9	1.7	0.8	2.1	1.1	1.2	1.7	1.3	1.0	1.9	1.7
S	1.1	1.3	1.2	1.4	1.1	1.1	2.0	0.6	1.0	1.0	1.1	1.7	2.2	1.8	1.8	1.6	1.4	0.7	4.5	1.5
T	0.7	0.8	0.9	1.0	0.5	0.5	1.4	1.1	1.0	0.6	0.5	1.1	2.1	1.4	0.9	1.0	0.5	0.5	7.3	0.8
V	0.3	0.0	0.8	1.0	0.4	0.2	0.8	0.1	0.9	0.3	0.4	0.9	1.6	1.8	0.7	0.8	0.4	0.1	0.5	0.4
W	0.6	0.6	2.5	4.0	0.8	9.9	8.0	0.7	2.9	0.5	0.9	10.4	1.1	6.9	4.7	7.0	2.4	0.1	0.0	1.4
Y	0.8	0.0	0.4	0.9	0.3	0.5	1.0	1.3	0.7	1.0	1.0	0.6	1.5	1.3	0.4	0.9	0.4	2.3	0.0	0.3

E

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	1.1	0.8	0.8	0.9	0.6	0.9	1.2	0.7	0.9	0.7	1.0	0.9	2.9	2.0	1.8	1.9	1.2	0.6	0.4	0.7
C	0.6	0.0	0.2	0.8	0.0	0.3	0.4	0.0	0.3	0.7	0.0	0.3	1.8	0.8	0.5	0.6	0.7	0.2	0.7	1.0
D	0.9	0.2	0.6	0.8	0.5	0.4	1.9	1.0	0.8	0.7	0.7	0.6	1.0	0.8	1.3	1.6	1.2	0.7	4.3	1.1
E	1.0	0.2	0.8	0.5	0.6	0.7	1.1	0.6	0.8	1.1	0.6	0.9	1.5	1.1	1.1	1.6	0.9	0.5	1.0	0.6
F	0.6	0.4	0.5	0.5	0.3	0.2	0.4	0.3	0.6	0.4	0.3	0.4	0.9	1.1	0.9	0.6	0.3	0.2	1.1	0.3
G	1.0	0.8	0.4	0.7	0.2	0.8	1.2	0.4	0.4	0.5	0.2	0.4	1.1	1.3	1.0	0.7	0.9	0.8	0.2	0.4
H	2.1	0.4	0.8	1.5	1.3	0.8	1.3	1.5	0.8	1.3	1.5	0.9	2.1	2.0	1.7	2.0	0.6	1.1	1.7	0.8
I	0.6	0.7	0.6	0.7	0.3	0.3	0.8	0.3	0.7	0.4	0.3	1.0	2.2	0.9	0.6	0.9	0.5	0.7	0.9	0.3
K	0.7	1.0	0.9	0.6	0.2	0.4	0.5	0.7	0.9	0.6	0.6	0.6	3.1	0.9	1.3	0.9	1.2	1.0	1.6	0.7
L	0.8	0.3	0.8	0.6	0.0	0.4	0.8	0.3	0.7	0.6	0.5	1.3	3.2	0.8	1.2	1.5	0.7	0.4	1.2	0.3
M	0.6	0.6	1.0	0.4	0.2	0.7	1.0	0.4	0.8	0.6	0.2	0.6	1.0	0.7	0.9	0.6	0.7	0.1	0.0	0.0
N	0.7	0.0	0.4	0.5	0.6	0.4	1.4	1.2	1.1	0.9	0.7	0.9	1.7	0.8	0.6	1.1	1.1	0.4	5.0	1.2
P	2.4	0.0	3.6	2.2	0.8	1.6	3.1	1.4	1.5	2.0	2.3	2.1	4.6	2.5	3.6	3.6	2.0	1.8	0.8	1.2
Q	1.3	0.3	1.3	1.0	0.9	0.9	2.0	0.7	1.6	1.0	1.5	1.8	2.5	1.7	1.9	1.5	2.2	1.2	5.0	1.1
R	1.6	1.1	0.9	1.0	1.1	2.3	1.8	0.7	1.5	1.3	0.6	1.1	2.9	1.6	1.9	1.7	1.4	1.1	0.6	0.8
S	1.7	1.2	1.2	1.2	0.6	1.1	2.5	0.8	1.4	1.1	0.7	0.9	3.3	2.4	2.3	1.9	1.4	0.8	0.8	0.8
T	1.3	0.2	0.9	0.8	0.7	0.4	1.6	1.0	0.7	1.2	0.5	1.1	2.0	1.3	1.7	1.1	1.6	0.6	1.6	0.2
V	0.8	0.3	0.4	0.8	0.3	0.3	0.8	0.6	0.9	0.6	0.5	0.3	1.7	1.5	1.2	0.9	0.6	0.5	0.2	0.1
W	1.0	0.7	1.4	1.6	0.4	0.7	0.9	0.8	1.3	0.3	2.4	0.7	0.6	3.2	2.1	2.6	0.8	0.5	0.0	0.3
Y	0.9	0.6	0.4	0.5	0.2	0.3	1.5	0.2	0.6	0.5	0.3	0.4	2.0	1.7	0.7	0.8	0.2	0.3	0.0	0.0

Fig. 4. Dipeptide propensity for the 41–60 residue length linker set. Interpretation of figure refers to legend to Fig. 1.

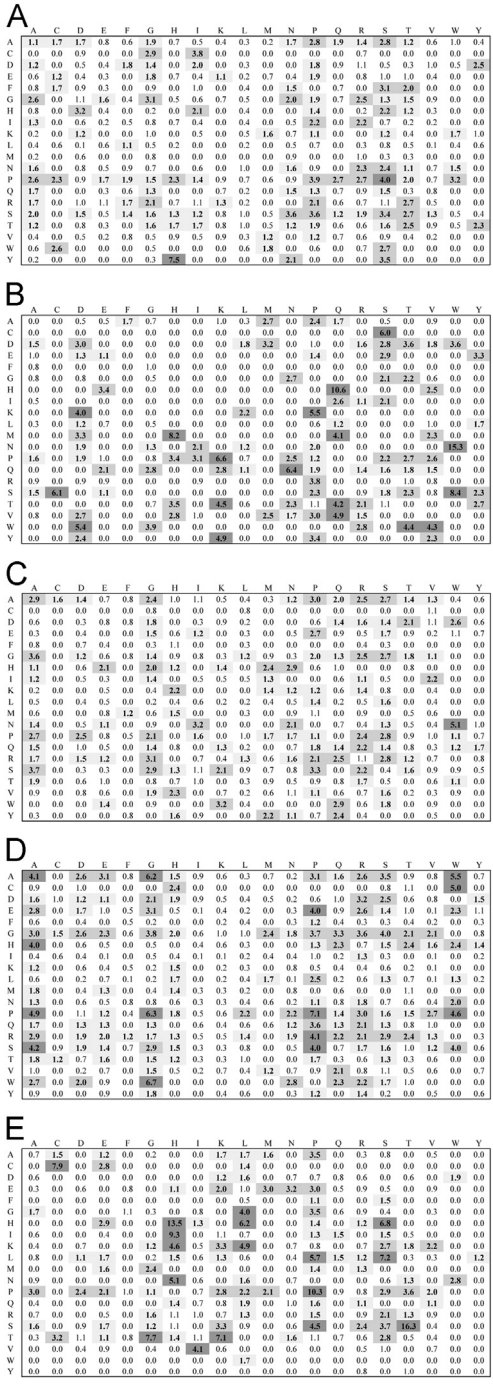

Fig. 5. Dipeptide propensity for the 61–200 residue length linker set. Interpretation of figure refers to legend to Fig. 1.

Table 7
Fraction of predicted secondary structure in linker regions.

	Secondary structure		
	α-helix	**β-strand**	**coil**
Actinobacteria	0.14	0.02	0.86
Alphaproteobacteria	0.19	0.03	0.78
Betaproteobacteria	0.30	0.01	0.69
Firmicutes	0.02	0.06	0.92
Gammaproteobacteria	0.26	0.02	0.72

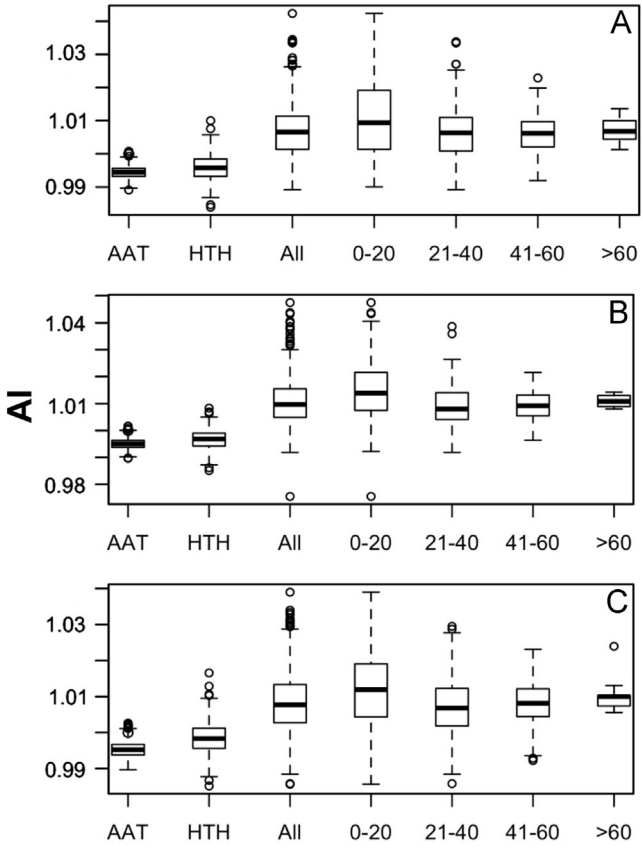

Fig. 6. Box-plots of the distribution of the average linker flexibility (index #425 of Table 2 in [1] and code VINM940101 in AAindex [10]). Horizontal axis indicates the average flexibility distribution in the wHTH, AAT domains, in all linkers, and in linkers belonging to different length intervals: 0–20, 21–40, 41–60 and >60 residues. Y-axis reports the flexibility scale (label AI stands for Average Index). A, B, and C, denote Alphaproteobacteria, Betaproteobacteria, and Gammaproteobacteria, respectively.

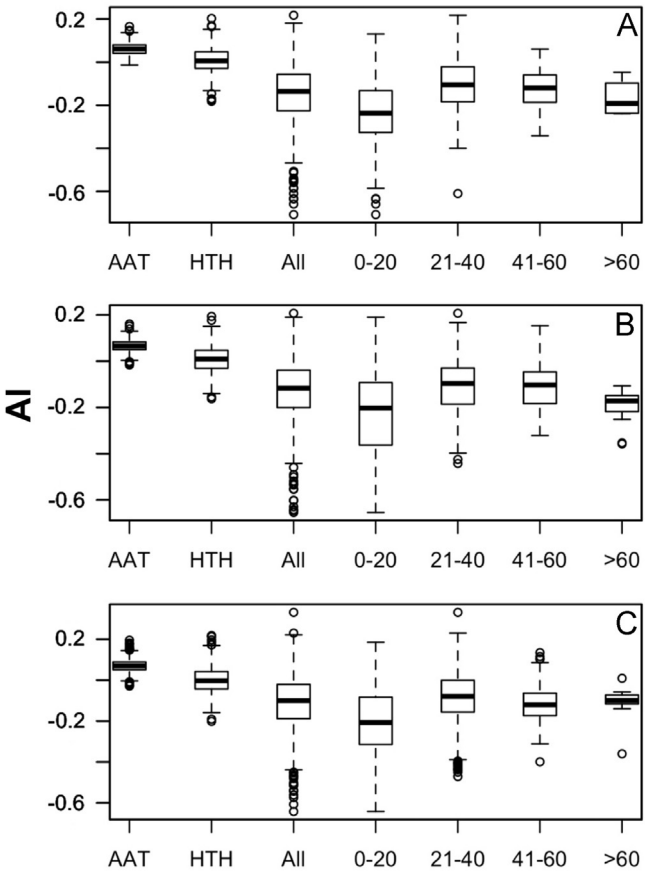

Fig. 7. Box plots of the distribution of average linker hydrophobicity (index #58 of Table 2 in [1] and code CIDH920105 in AAindex [10]). For interpretation of plots, refer to Fig. 6 caption.

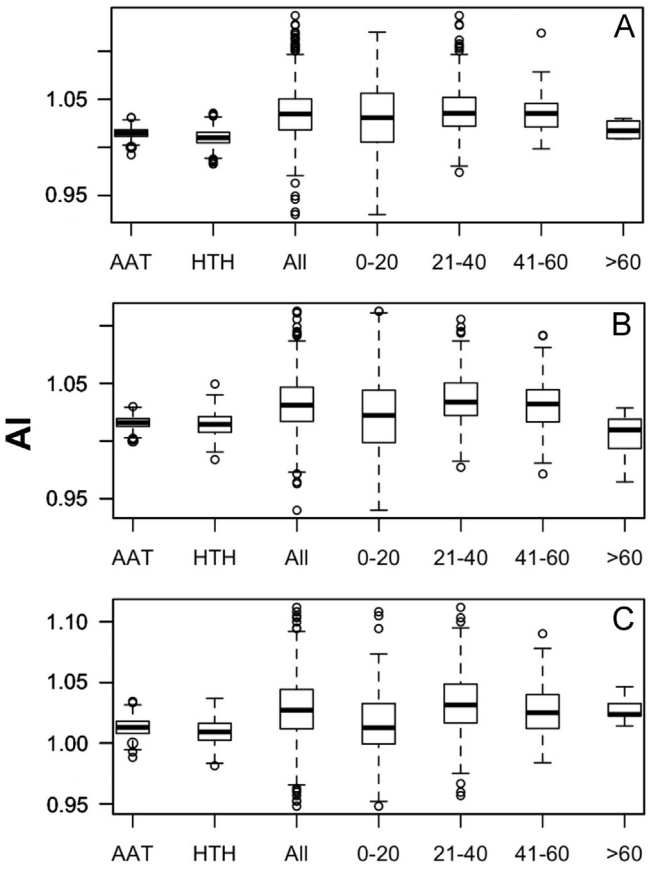

Fig. 8. Box plots of the distribution of average Linker propensity index (#491 of Table 2 in [1] and code GEOR03010 in AAindex [10]). For interpretation of plots, refer to Fig. 6 caption.

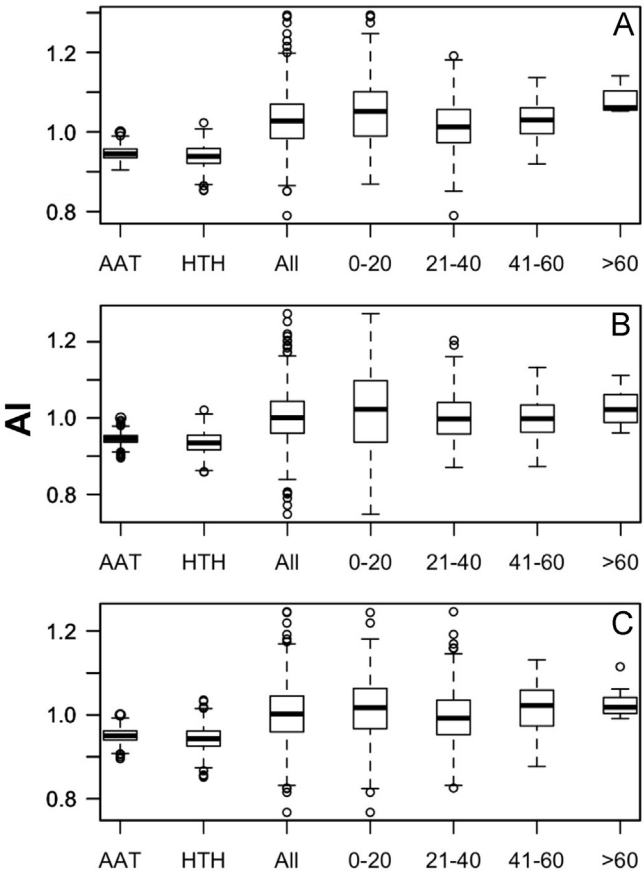

Fig. 9. Box plots of the distribution of the average normalized β-turn propensity (index #37 Table 2 in [1] and code CHOP780101 in AAindex [10]). For interpretation of plots, refer to Fig. 6 caption.

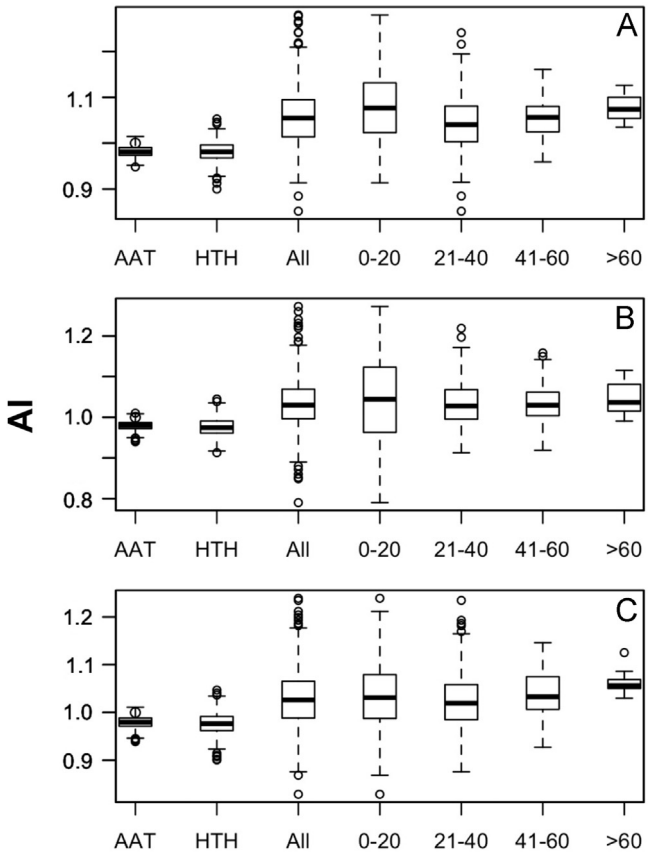

Fig. 10. Box plots of the distribution of the average Chou–Fasman coil propensity (#24 of Table 2 in [1] and code CHAM830101 in AAindex [10]). For interpretation of plots, refer to Fig. 6 caption.

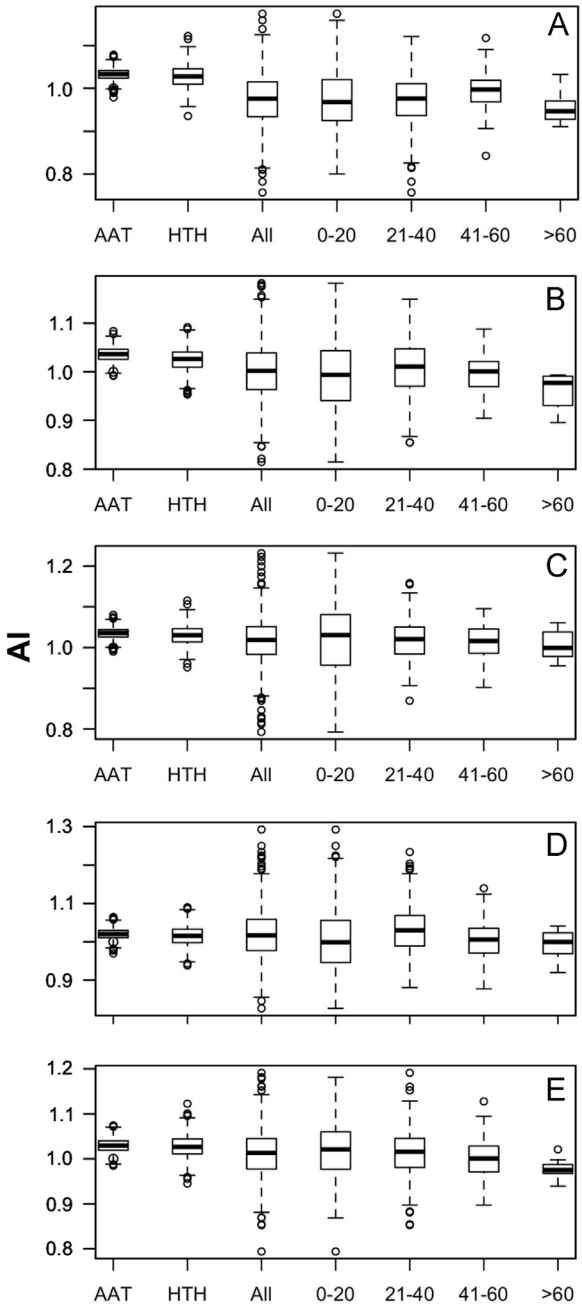

Fig. 11. Box plots of the distribution of average normalized α-helix propensity (index #38 of Table 2 in [1] and code CHOP780102 in AAindex [10]). A, B, C, D and E denote Actinobacteria, Alphaproteobacteria, Betaproteobacteria, Firmicutes and Gammaproteobacteria, respectively.

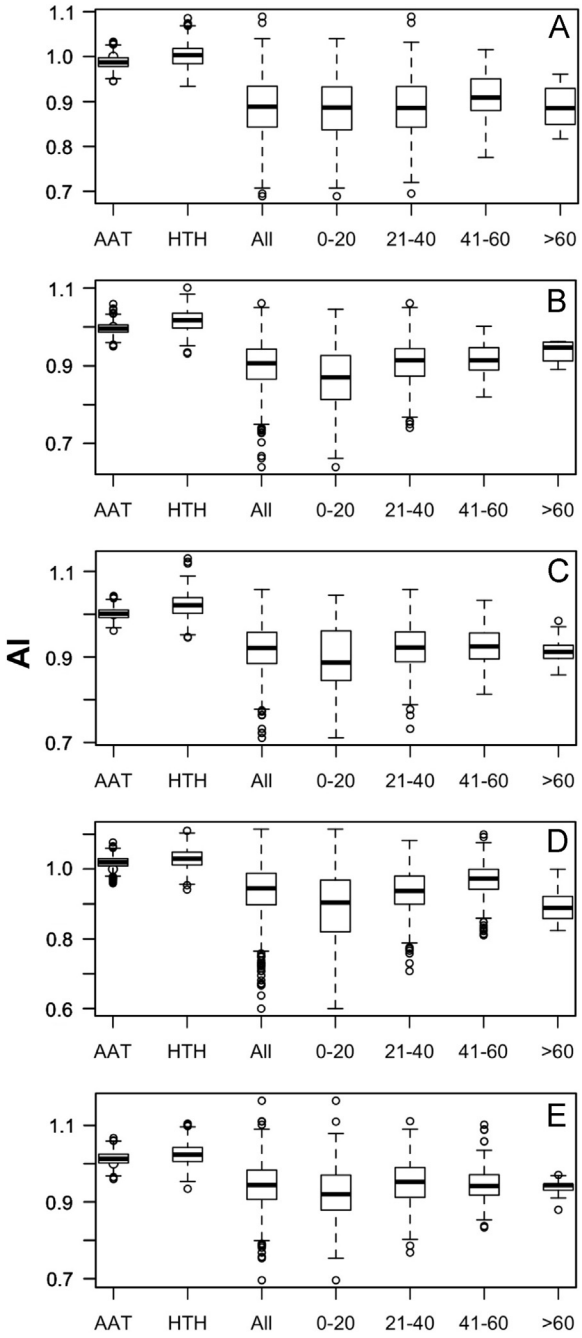

Fig. 12. Box plots of the distribution of average normalized β-sheet propensity (index #39 of Table 2 in [1] and code CHOP780103 in AAindex [10]). Letter interpretation is as in Fig. 11 caption.

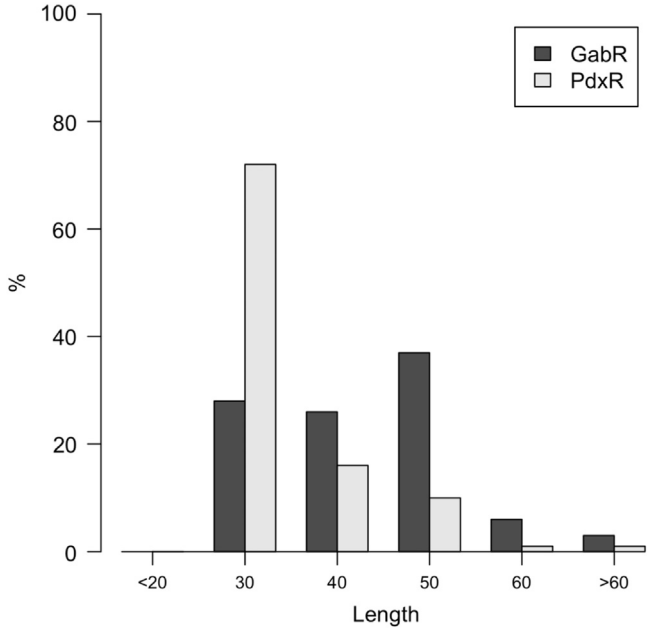

Fig. 13. Histogram of the linker length distribution in the MocR subgroups GabR and PdxR. Horizontal axis labels indicate length intervals: 20 corresponds to 0–20, 30 (21–30), 40 (31–40), 50 (41–50), 60 (51–60) and >60 (longer than 60 residues). Percentage (%) on the vertical axis indicates the fraction of linkers in the length interval. Sequences were retrieved from the reference proteomes data bank available at the Hmmer web server [17] using a significance E-value thresholds equal to 10^{-120}. With this threshold, 885 and 334 sequences were retrieved for GabR and PdxR, respectively.

Table 8
GabR and PdxR sequences retrieved from RegPrecise data bank.

GabR		
UniProt code	**Specie**	**Phylum**
A0A098SFD5	*Acinetobacter baumannii* AB0057	*Gammaproteobacteria*
Q6F766	*Acinetobacter sp.* AD	*Gammaproteobacteria*
A7Z1D7	*Bacillus amyloliquefaciens* FZB42	*Firmicutes*
A8F9Y9	*Bacillus pumilus* SAFR 032	*Firmicutes*
P94426	*Bacillus subtilis subsp. subtilis str.* 168	*Firmicutes*
Q2KX56	*Bordetella avium* 197N	*Betaproteobacteria*
A0A0H3LKN1	*Bordetella bronchiseptica* RB50	*Betaproteobacteria*
Q0B6G3	*Burkholderia cepacia* AMMD	*Betaproteobacteria*
C5ALU9	*Burkholderia glumae* BGR1	*Betaproteobacteria*
A0A0H2XDM4	*Burkholderia mallei* ATCC 23344	*Betaproteobacteria*
B2JSD8	*Burkholderia phymatum* STM815	*Betaproteobacteria*
B2JR38	*Burkholderia phymatum* STM815	*Betaproteobacteria*
Q63NL7	*Burkholderia pseudomallei* K96243	*Betaproteobacteria*
A4JJX2	*Burkholderia vietnamiensis* G4	*Betaproteobacteria*
Q13LC0	*Burkholderia xenovorans* LB400	*Betaproteobacteria*
A9BMY2	*Delfia acidovorans* SPH-1	*Betaproteobacteria*
D4HXE9	*Erwinia amylovora* ATCC 49946	*Gammaproteobacteria*
Q6D5I8	*Erwinia carotovora subsp.atroseptica* SCRI1043	*Gammaproteobacteria*
A6TF79	*Klebsiella pneumonia subsp. pneumoniae* MGH 78578	*Gammaproteobacteria*
B2U7Y5	*Ralstonia pickettii* 12J	*Betaproteobacteria*
A8GJW1	*Serratia proteamaculans* 568	*Gammaproteobacteria*
Q4A0R1	*Staphylococcus saprophyticus subsp. saprophyticus* ATCC 15305	*Firmicutes*
C4ZIR5	*Thauera sp.*MZ1T	*Betaproteobacteria*
Q7CJK7	*Yersinia pestis* KIM	*Gammaproteobacteria*
A1VQK3	*Polaromonas naphthalenivorans* CJ2	*Betaproteobacteria*
Q129G7	*Polaromonas sp.* JS666	*Betaproteobacteria*
Q221G1	*Rhodoferax ferrireducens* DSM 15236	*Betaproteobacteria*
C5CM40	*Variovorax paradoxus* S110	*Betaproteobacteria*
PdxR		
B9MKZ0	*Anaerocellum thermophilum* DSM6725	*Firmicutes*
A4XIB4	*Caldicellulosiruptor saccharolyticus* DSM 8903	*Firmicutes*
Q929S0	*Listeria innocua* Clip11262	*Firmicutes*
Q8Y5G3	*Listeria monocytogenes* EGD e	*Firmicutes*
A0AKK7	*Listeria welshimeri serovar 6b str.* SLCC5334	*Firmicutes*
C7MF20	*Brachybacterium faecium* DSM 4810	*Actinobacteria*
Q6AFC0	*Leifsonia xyli subsp.xyli str.* CTCB07	*Actinobacteria*
B3GXB5	*Actinobacillus pleuropneumoniae servar 7 str.* AP76	*Gammaproteobacteria*
Q5WKW3	*Bacillus clausii* KSM K16	*Firmicutes*
C3PLB2	*Corynebacterium aurimucosum* ATCC 700975	*Actinobacteria*
Q6NK11	*Corynebacterium diphtheriae* NCTC 13129	*Actinobacteria*
Q8NS92	*Corynebacterium glutamicum* ATCC 13032	*Actinobacteria*
B2GK63	*Kocuria rhizophila* DC2201	*Actinobacteria*
B9E8T3	*Macrococcus caseolyticus* JCSC5402	*Firmicutes*
W8TRW2	*Staphylococcus aureus subsp. aureus* N325	*Firmicutes*
B9DKX6	*Staphylococcus aureus subsp.carnosus* TM300	*Firmicutes*

Table 8 (*continued*)

GabR		
UniProt code	**Specie**	**Phylum**
A0A0H2VKR4	*Staphylococcus epidermidis* ATCC 12228	*Firmicutes*
A0A0Q1AKJ7	*Staphylococcus haemolyticus* JCSC1435	*Firmicutes*
Q49V27	*Staphylococcus saprophyticus subsp. saprophyticus* ATCC15035	*Firmicutes*

Acknowledgements

This work was supported by Regione Lazio [grant code FILAS-RU-2014-1020 A/15/2015 to TM], by University of Rome "La Sapienza", Italy and by the Italian Education, University and Research Ministry, Italy (MIUR) [grant numbers C26N158EP9; C26A14SY4E].

Transparency document. Supplementary material

Transparency data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.dib.2016.08.064.

References

[1] T. Milano, S. Angelaccio, A. Tramonti, M.L. Di Salvo, R. Contestabile, S. Pascarella, Structural properties of the linkers connecting the N- and C- terminal domains in the MocR bacterial transcriptional regulators, Biochim. Open 3 (2016) 8–18.
[2] T. Tatusova, S. Ciufo, B. Fedorov, K. O'Neill, I. Tolstoy, RefSeq microbial genomes database: new representation and annotation strategy, Nucleic Acids Res. 42 (2014) D553–D559.
[3] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[4] A. Marchler-Bauer, M.K. Derbyshire, N.R. Gonzales, S. Lu, F. Chitsaz, L.Y. Geer, R.C. Geer, J. He, M. Gwadz, D.I. Hurwitz, C. J. Lanczycki, F. Lu, G.H. Marchler, J.S. Song, N. Thanki, Z. Wang, R.A. Yamashita, D. Zhang, C. Zheng, S.H. Bryant, CDD: NCBI's conserved domain database, Nucleic Acids Res. 43 (2015) D222–D226.
[5] Y. Huang, B. Niu, Y. Gao, L. Fu, W. Li, CD-HIT Suite: a web server for clustering and comparing biological sequences, Bioinformatics 26 (2010) 680–682.
[6] F. Sievers, A. Wilm, D. Dineen, T.J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Soding, J.D. Thompson, D. G. Higgins, Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega, Mol. Syst. Biol. 7 (2011) 539.
[7] A.M. Waterhouse, J.B. Procter, D.M. Martin, M. Clamp, G.J. Barton, Jalview Version 2–a multiple sequence alignment editor and analysis workbench, Bioinformatics 25 (2009) 1189–1191.
[8] P. Rice, I. Longden, A. Bleasby, EMBOSS: the European Molecular Biology Open Software Suite, Trends Genet. 16 (2000) 276–277.
[9] D. Frishman, P. Argos, Seventy-five percent accuracy in protein secondary structure prediction, Proteins 27 (1997) 329–335.
[10] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, M. Kanehisa, AAindex: amino acid index database, progress report 2008, Nucleic Acids Res. 36 (2008) D202–D205.
[11] D. Heider, Interpol: Interpolation of amino acid sequences, R package version 1.3.1. (2012).
[12] R Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 2016.
[13] B.R. Belitsky, Bacillus subtilis GabR, a protein with DNA-binding and aminotransferase domains, is a PLP-dependent transcriptional regulator, J. Mol. Biol. 340 (2004) 655–664.
[14] A. Tramonti, A. Fiascarelli, T. Milano, M.L. di Salvo, I. Nogues, S. Pascarella, R. Contestabile, Molecular mechanism of PdxR - a transcriptional activator involved in the regulation of vitamin B6 biosynthesis in the probiotic bacterium *Bacillus clausii*, FEBS J. 282 (2015) 2966–2984.
[15] P.S. Novichkov, A.E. Kazakov, D.A. Ravcheev, S.A. Leyn, G.Y. Kovaleva, R.A. Sutormin, M.D. Kazanov, W. Riehl, A.P. Arkin, I. Dubchak, D.A. Rodionov, RegPrecise 3.0–a resource for genome-scale exploration of transcriptional regulation in bacteria, BMC Genom. 14 (2013) 745.
[16] S.R. Eddy, Profile hidden Markov models, Bioinformatics 14 (1998) 755–763.
[17] R.D. Finn, J. Clements, W. Arndt, B.L. Miller, T.J. Wheeler, F. Schreiber, A. Bateman, S.R. Eddy, HMMER web server: 2015 update, Nucleic Acids Res. 43 (2015) W30–W38.