Supplementary Material

Convergent behavior of extended stalk regions from staphylococcal surface proteins with widely divergent sequence patterns

Alexander E. Yarawsky[1#], Andrea L. Ori[1,2†], Lance R. English[3‡], Steven T. Whitten[3], and Andrew B. Herr[1,4,5]

Affiliations:
[1] Division of Immunobiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, USA
[2] Medical Sciences Baccalaureate Program, University of Cincinnati, Cincinnati, OH 45267, USA
[3] Department of Chemistry and Biochemistry, Texas State University, San Marcos, TX 78666, USA
[4] Division of Infectious Diseases, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, USA
[5] Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH 45229, USA

[#] Current affiliation: BioAnalysis, LLC, Philadelphia, PA 19134, USA
[†] Current affiliation: Graduate Program in Molecular Biophysics, Johns Hopkins University, Baltimore, MD 21218, USA
[‡] Current affiliation: Department of Physical Sciences, Temple College, Temple, TX 76502, USA

Correspondence to Andrew B. Herr: Division of Immunobiology, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, Cincinnati, OH 45229, USA. andrew.herr@cchmc.org

Supplementary data files present in this document include:

1. **Supplementary Figure S1**
2. **Supplementary Tables S1—S5**

Supplementary Figure

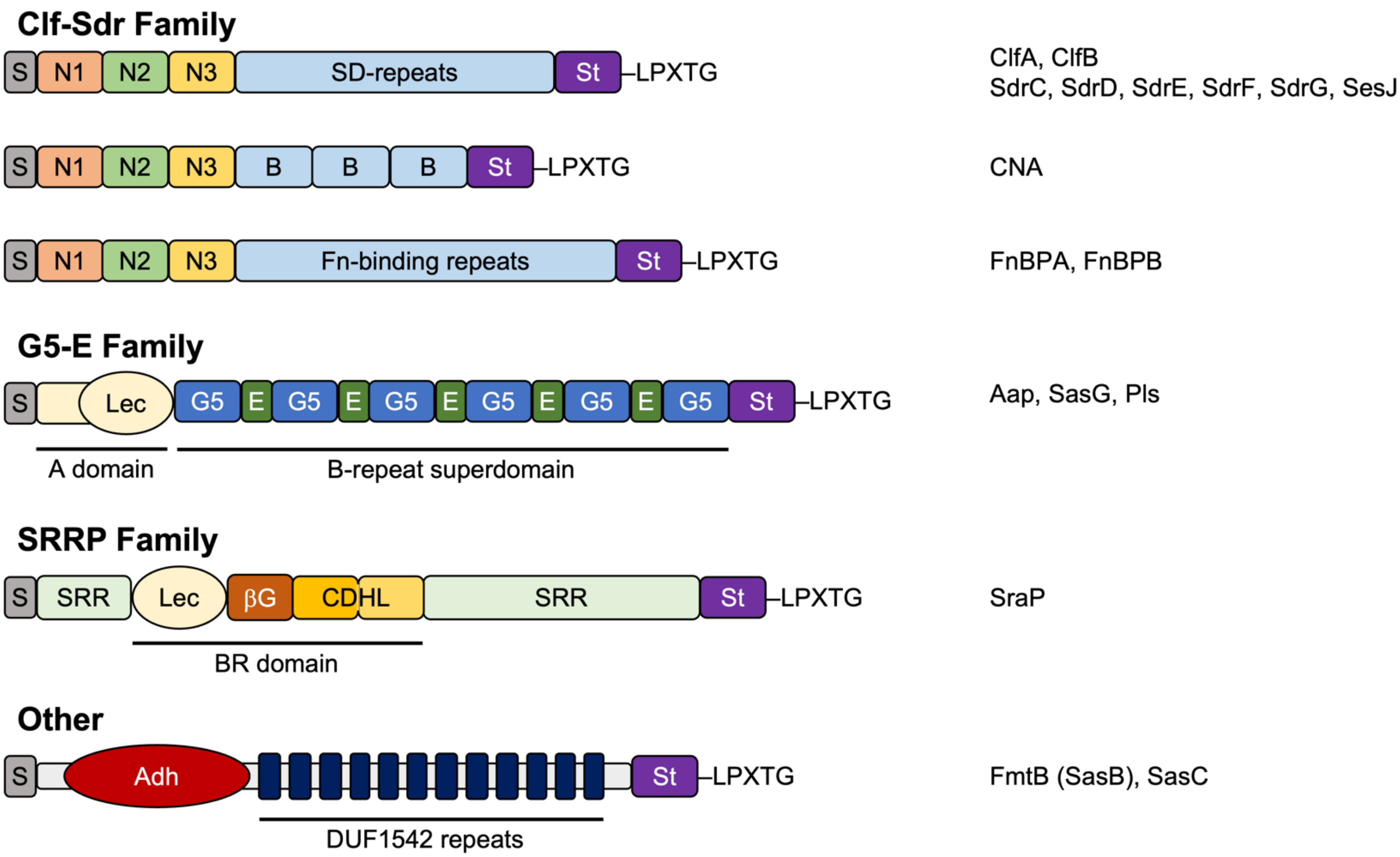

Figure S1. Domain organization of adhesin-like CWA protein families.
The major families and sub-families of staphylococcal adhesin-like CWA proteins are illustrated with the representative proteins from each family listed on the right. The gray S box represents the signal sequence. The purple St box is the stalk region, although the nature of this region varies per protein and may have distinct characteristics even within a family, as defined in Table 3. For example, CNA is in the Clf-Sdr family but it has a Pro-rich stalk region; likewise, Pls is in the G5-E family but is has a SD-rich stalk, unlike Aap and SasG. The LPXTG motif at the C-terminus of each protein is the sortase anchor sequence that is covalently attached to the cell wall of the staphylococcal cell. In the Clf-Sdr family, N1, N2, and N3 domains together form the A region; the N2 and N3 domains adopt Ig-like folds that interact with ligands via the 'dock, lock, and latch' mechanism. Lec stands for the lectin domains of Aap, SasG, Pls, and SraP. The B-repeat superdomain of Aap, SasG, and Pls is made up of tandem B-repeats, each of which comprises a G5 and E subdomain. The SRRP family (serine-rich repeat proteins) such as SraP contain serine-rich repeats (SRR), a lectin domain, a β-grasp (βG) fold domain, and two cadherin-like (CDHL) domains. The Other family contains FmtB (SasB) and SasC, which share a distinct domain arrangement with an adhesion domain followed by several DUF1542 repeats.

Supplementary Tables

Table S1. Sequence parameters for IDP constructs.

Parameter	Aap-PGR	SasG-PGR	Aap-Arpts	SdrC-SD	SD-30mer
N	135	69	189	62	30
f-	0.15556	0.13043	0.22222	0.33871	0.50000
f+	0.1037	0.23188	0.06878	0.08065	0
FCR	0.25926	0.36232	0.29101	0.41935	0.50000
NCPR	-0.05185	0.10145	-0.15344	-0.25806	-0.50000
Kappa	0.05825	0.09562	0.08655	0.30207	0.02324
SCD	0.79	1.11	15.80	4.06	10.79
FPR	0.28889	0.17391	0.09524	0.03226	0
Omega	0.03234	0.07106	0.05146	0.00657	0.00096
Hydropathy	3.09259	2.72899	3.08466	2.64839	2.35
Phase Plot Region	2	3	2	3	4

The CIDER server [1] was used to calculate most parameters, including those required for the Das-Pappu Plot [2]; SCD was calculated as described [3].
N: Number of residues
f_-: Fraction of negative residues
f_+: Fraction of positive residues
FCR: Fraction of charged residues
NCPR: Net charge per residue
Kappa: κ is a charge patterning parameter [2]. Highly mixed charged sequences approach κ = 0, while highly segregated charged sequences approach κ = 1.
SCD: Sequence charge decoration [3]. Well-mixed charged sequences approach SCD = 0, whereas highly segregated charged sequences show large values of SCD.
FPR: Fraction of proline residues (not a parameter provided by CIDER, but included here for relevance)
Omega: Ω is a charge/proline patterning parameter [4]. This parameter is similar to κ, but also incorporates proline residues. If prolines and charged residues are well mixed along a sequence (with respect to other amino acids), there will be a low Ω value. If proline/charged residues are highly segregated, Ω will approach 1.
Hydropathy: Based on the Kyte-Doolittle scale [5], normalized from 0 (least hydrophobic) to 9 (most hydrophobic).
Phase Plot Region: Location on the Das-Pappu phase plot this sequence falls
Phase Plot Annotation:
1: Weak polyampholytes and polyelectrolytes (Globules & Tadpoles)
2: Boundary region (Janus sequences)
3: Strong polyampholytes
4: Strong negatively charged polyelectrolytes
5: Strong positively charged polyelectrolytes

Table S2. Calculated and predicted parameters of IDP constructs.

IDP	N	Net charge	R_h (coil)	R_h (PPII)	R_h (PPII charge)	f_{PPII}
Aap-PGR	135	-7	25.64	38.50	37.84	0.5350
SasG-PGR	69	+7	18.27	24.56	24.43	0.4761
Aap-Arpts	189	-29	30.38	41.26	44.06	0.4190
SdrC-SD	62	-16	17.31	20.64	22.15	0.3294
SD-30mer	30	-15	12.01	13.45	15.16	0.2700

The number of residues is listed in the N column. R_h is the predicted hydrodynamic radius (in Å) assuming complete random coil (R_h (coil)), considering intrinsic propensities for the polyproline type-II helix backbone conformation (R_h (PPII)) or contributions from both PPII propensity and the net charge (R_h (PPII charge)). The predicted fraction of PPII (f_{PPII}) refers to the number of residues predicted to be in the PPII conformation divided by the total number of residues. All parameters were calculated using a program based on Tomasso, et al. [6]. Net charge contributions to the R_h were established empirically in English, et al. [7]

Table S3. Sequence-based parameters of IDP dataset. The dataset is reproduced from Tomasso, et al. [6]. Parameters listed here were calculated using a program provided by Steven Whitten, based on Tomasso, et al. [6]. Shaded IDPs are from the current study. IDPs are sorted by descending f_{PPII}.

IDP	N	Net charge	R_h (coil)	R_h (PPII)	R_h (PPII charge)	R_h[a] (experimental)	f_{PPII}
Aap-PGR	135	-7	25.64	38.50	37.84	37.06	0.5350
p53(1-93)	93	-15	21.24	29.51	30.56	32.4	0.4890
SasG-PGR	69	+7	18.27	24.56	24.43	24.8	0.4761
p53(1-93) ALA-	93	-15	21.24	28.66	29.70	30.4	0.4581
p53 TAD	73	-14	18.80	24.79	25.84	23.8	0.4500
Aap-Arpts	189	-29	30.38	41.26	44.06	40.8	0.4190
Securin	202	-1	31.41	42.57	40.45	39.7	0.4130
PDE-γ	87	+4	20.54	26.51	25.70	24.8	0.4122
Cad136	136	+9	25.73	33.77	33.45	28.1	0.4025
HIF1-α-403	202	-29	31.41	42.13	44.86	44.3	0.4024
Tau-K45	198	+19	31.10	41.52	42.53	45	0.3988
HIF1-α-530	170	-10	28.80	37.81	37.44	38.3	0.3899
Fos-AD	168	-16	28.62	37.17	37.84	35	0.3783
ShB-C	146	-4	26.67	34.32	33.06	32.9	0.3764
α-synuclein	140	-9	26.11	33.47	33.12	28.2	0.3744
Mlph(147-403)	260	-28	35.68	47.00	49.24	49	0.3703
CFTR-R-region	189	-5	30.38	39.18	37.82	32	0.3644
p57-ID	73	-6	18.80	23.14	22.80	24	0.3636
prothymosin-α	110	-43	23.12	29.02	34.77	33.7	0.3633
LJIDP1	94	+4	21.36	26.46	25.59	24.52	0.3565
Mlph(147-240)	97	-15	21.70	26.85	27.86	28	0.3528
SNAP25	206	-14	31.73	40.60	40.70	39.7	0.3513
Hdm2-ABD	97	-29	21.70	26.47	29.91	25.7	0.3345
SdrC-SD	62	-16	17.31	20.64	22.15	21.1	0.3294
Vmw65	89	-19	20.78	25.13	26.90	28	0.3278
p53(1-93) PRO-	93	-15	21.24	24.93	25.97	27.4	0.2832
SD-30mer	30	-15	12.01	13.45	15.16	ND[b]	0.2700

[a] Reported in Å. Values in gray cells were as determined in this manuscript or [8]; values in white cells are reproduced from [6].
[b] ND, not determined.

Table S4. The sequence of IDPs used in PPII and R_h predictions. IDP sequences (other than those from the current study - shaded) are from Tomasso, et al. supplementary material [6].

IDP	Sequence
p53(1-93)	MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLS PDDIEQWFTEDPGPDEAPRMPEAAPPVAPAPAAPTPAAPAPAPSW PL
p53(1-93) ALA-	MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQGMDDLMLS PDDIEQWFTEDPGPDEGPRMPEGGPPVGPGPGGPTPGGPGPGPS WPL
p53(1-93) PRO-	MEEGQSDGSVEGGLSQETFSDLWKLLGENNVLSGLGSQAMDDLML SGDDIEQWFTEDGGGDEAGRMGEAAGGVAGAGAAGTGAAGAGAG SWGL
p53 TAD	MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLS PDDIEQWFTEDPGPDEAPRMPEAAPRV
Vmw65	GSAGHTRRLSTAPPTDVSLGDELHLDGEDVAMAHADALDDFDLDML GDGDSPGPGFTPHDSAPYGALDMADFEFEQMFTDALGIDEYGG
Hdm2-ABD	ERSSSSESTGTPSNPDLDAGVSEHSGDWLDQDSVSDQFSVEFEVE SLDSEDYSLSEEGQELSDEDDEVYQVTVYQAGESDTDSFEEDPEIS LADYWK
prothymosin-α	MSDAAVDTSSEITTKDLKEKKEVVEEAENGRDAPANGNANEENGEQ EADNEVDEEEEEGGEEEEEEEEGDGEEEDGDEDEEAESATGKRAA EDDEDDDVDTKKQKTDEDD
HIF1-α-403	PAAGDTIISLDFGSNDTETDDQQLEEVPLYNDVMLPSPNEKLQNINLA MSPLPTAETPKPLRSSADPALNQEVALKLEPNPESLELSFTMPQIQD QTPSPSDGSTRQSSPEPNSPSEYCFYVDSDMVNEFKLELVEKLFAE DTEAKNPFSTQDTDLDLEMLAPYIPMDDDFQLRSFDQLSPLESSSAS PESASPQSTVTVFQ
Fos-AD	GSHMSVASLDLTGGLPEVATPESEEAFTLPLLNDPEPKPSVEPVKSI SSMELKTEPFDDFLFPASSRPSGSETARSVPDMDLSGSFYAADWEP LHSGSLGMGPMATELEPLCTPVVTCTPSCTAYTSSFVFTYPEADSFP SCAAAHRKGSSSNEPSSDSLSSPTLLAL
Mlph(147-240)	RLQGGGGSEPSLEEGNGDSEQTDEDGDLDTEARDQPLNSKKKKRL LSFRDVDFEEDSDHLVQPCSQTLGLSSVPESAHSLQSLSGEPYSED TTSLEP
Tau-K45	MSSPGSPGTPGSRSRTPSLPTPPTREPKKVAVVRTPPKSPSSAKSR LQTAPVPMPDLKNVKSKIGSTENLKHQPGGGKVQIINKKLDLSNVQS KCGSKDNIKHVPGGGSVQIVYKPVDLSKVTSKCGSLGNIHHKPGGG QVEVKSEKLDFKDRVQSKIGSLDNITHVPGGGNKKIETHKLTFRENA KAKTDHGAEIVY
Mlph(147-403)	RLQGGGGSEPSLEEGNGDSEQTDEDGDLDTEARDQPLNSKKKKRL LSFRDVDFEEDSDHLVQPCSQTLGLSSVPESAHSLQSLSGEPYSED TTSLEPEGLEETGARALGCRPSPEVQPCSPLPSGEDAHAELDSPAA SCKSAFGTTAMPGTDDVRGKHLPSQYLADVDTSDEDSIQGPRAASQ HSKRRARTVPETQILELNKRMSAVEHLLVHLENTVLPPSAQEPTVET HPSADTEEETLRRRLEELTSNISGSSTSSE
p57-ID	VRTSACRSLFGPVDHEELSRELQARLAELNAEDQNRWDYDFQQDM PLRGPGRLQWTEVDSDSVPAFYRETVQV
PDE-γ	MNLEPPKAEIRSATRVMGGPVTPRKGPPKFKQRQTRQFKSKPPKK GVQGFGDDIPGMEGLGTDITVICPWEAFNHLELHELAQYGII

LJIDP1	MARSFTNIKAISALVAEEFSNSLARRGYAATAQSAGRVGASMSGKM GSTKSGEEKAAAREKVSWVPDPVTGYYKPENIKEIDVAELRSAVLGK N
Cad136	RLEQYTSAVVGNKAAKPAKPAASDLPVPAEGVRNIKSMWEKGNVFS SPGGTGTPNKETAGLKVGVSSRINEWLTKTPEGNKSPAPKPSDLRP GDVSGKRNLWEKQSVEKPAASSSKVTATGKKSETNGLRQFEKEP
α-synuclein	MDVFMKGLSKAKEGVVAAAEKTKQGVAEAAGKTKEGVLYVGSKTK EGVVHGVATVAEKTKEQVTNVGGAVVTGVTAVAQKTVEGAGSIAAA TGFVKKDQLGKNEEGAPQEGILEDMPVDPDNEAYEMPSEEGYQDY EPEA
CFTR-R-region	GAMESAERRNSILTETLHRFSLEGDAPVSWTETKKQSFKQTGEFGE KRKNSILNPINSIRKFSIVQKTPLQMNGIEEDSDEPLERRLSLVPDSEQ GEAILPRISVISTGPTLQARRRQSVLNLMTHSVNQGQNIHRKTTASTR KVSLAPQANLTELDIYSRRLSQETGLEISEEINEEDLKECLFDDME
SNAP25	MAEDADMRNELEEMQRRADQLADESLESTRRMLQLVEESKDAGIR TLVMLDEQGEQLERIEEGMDQINKDMKEAEKNLTDLGKFCGLCVCP CNKLKSSDAYKKAWGNNQDGVVASQPARVVDEREQMAISGGFIRR VTNDARENEMDENLEQVSGIIGNLRHMALDMGNEIDTQNRQIDRIME KADSNKTRIDEANQRATKMLGSG
ShB-C	MTLGQHMKKSSLSESSSDMMDLDDGVESTPGLTETHPGRSAVAPF LGAQQQQQQPVASSLSMSIDKQLQHPLQQLTQTQLYQQQQQQQ QQQNGFKQQQQQTQQQLQQQQSHTINASAAAATSGSGSSGLTMR HNNALAVSIETDV
HIF1-α-530	NEFKLELVEKLFAEDTEAKNPFSTQDTDLDLEMLAPYIPMDDDFQLR SFDQLSPLESSSASPESASPQSTVTVFQQTQIQEPTANATTTTATTD ELKTVTKDRMEDIKILIASPSPTHIHKETTSATSSPYRDTQSRTASPNR AGKGVIEQTEKSHPRSPNVLSVALSQR
Securin	MATLIYVDKENGEPGTRVVAKDGLKLGSGPSIKALDGRSQVSTPRF GKTFDAPPALPKATRKALGTVNRATEKSVKTKGPLKQKQPSFSAKK MTEKTVKAKSSVPASDDAYPEIEKFFPFNPLDFESFDLPEEHQIAHLP LSGVPLMILDEERELEKLFQLGPPSPVKMPSPPWESNLLQSPSSILS TLDVELPPVCCDIDI
Aap-PGR	AEPGKPAEPGKPAEPGKPAEPGTPAEPGKPAEPGTPAEPGKPAEP GKPAEPGKPAEPGKPAEPGTPAEPGTPAEPGKPAEPGTPAEPGKP AEPGTPAEPGKPAESGKPVEPGTPAQSGAPEQPNRSMHSTDNKNQ
SasG-PGR	PKDPKGPENPEKPSRPTHPSGPVNPNNPGLSKDRAKPNGPVHSMD KNDKVKKSKIAKESVANQEKKRAE
Aap-Arpts	NNEAPQMSSTLQAEEGSNAEAPQSEPTKAEEGGNAEAAQSEPTKA EEGGNAEAPQSEPTKAEEGGNAEAAQSEPTKTEEGSNVKAAQSEP TKAEEGSNAEAPQSEPTKTEEGSNAKAAQSEPTKAEEGGNAEAAQ SEPTKTEEGSNAEAPQSEPTKAEEGGNAEAPQSEPTKTEEGGNAE APNVPTIKA
SdrC-SD	SDSDSDSDSDSDSDSDSDSDSDSDSDSDNDSDSDSDSDSDAGKHT PAKPMSTVKDQHKTAKA
SD-30mer	SDSDSDSDSDSDSDSDSDSDSDSDSDSDSD

Table S5: The sequence of low-complexity regions from Staphylococcal CWA proteins from Table 5. Sequences start at the beginning of the consensus LCR region identified by the PlaToLoCo server [9] and extend through the sequence immediately upstream of the LPXTG sortase motif. See Materials and Methods for further details.

Protein	Sequence
SD-rich LCRs	
SdrC	TSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSNSD SDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDS DSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSD NDSDSDSDSDSDAGKHTPAKPMSTVKDQHKTAKA
SdrD	TSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSD SDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDS DSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDAGKHTPVKPMSATKDH HNKAKA
SdrE	TSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSD SDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDS DSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSD AGKHTPVKPMSTTKDHHNKAKA
SdrF (*S. epi*)	TSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSD SDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDS DSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSD NDSDSDSDSDSDAGKHTPAKPMSTVKDQHKTAKA
SdrG (*S. epi*)	TSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSD SDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDS DSDSDSDSDSDSDSDSDSDSDSDNDSDSDSDSDSDAGKHTPAKPMSTVKDQHK TAKA
Pls	DSDADSDSDADSDSDADSDSDADSDSDADSDSDSDSDSDSDADSDSDSD SDSDADSDSDADSDSDADSDSDSDADSDSDSDSDSDADSDSDSDADSDSDADS DSDSDSDSDADSDSDSDSDSDADSDSDADSDSDADSDSDADSDSDSDSDSDAD SDSDADSDSDADSDSDADSDSDSDSDSDADSDSDSDSDSDSDADSDSDADSDS DSDADSDSDADSDSDADSDSDADSDSDSDSDSDADSDSDADSDSDADSDSDAD SDSDSDSDSDSDSDSDADSDSDSDSDSDADRDHNDKTDKPNNKE
ClfA	VPEQPDEPGEIEPIPEDSDSDPGSDSGSDSNSDSGSDSGSDSTSDSGSDSASDS DSASDSDSASDSDSASDSDSASDSDSDNDSDSDSDSDSDSDSDSDSDSDSDSD SDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDS DSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSD SDSDSDSDSDSASDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSES DSDSESDSDSDSDSDSDSDSDSDSDSDSASDSDSGSDSDSSSDSDSESDSNSD SESGSNNNVVPPNSPKNGTNASNKNEAKDSKEP
ClfB	VDPEPSPDPEPEPTPDPEPSPDPEPEPSPDPDPDSDSDSDSGSDSDSGSDSDSE SDSDSDSDSDSDSDSDSESDSDSESDSDSDSDSDSDSDSESDSDSDSDSDS DSDSDSESDSDSESDSESDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSD SESDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDS DSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSDSRVTPPNNEQKAPS NPKGEVNHSNKVSKQHKTDA
SesJ (*S. epi*)	FEDSESDSSSESESDSESHSDSESHSDSESTSESDSESHSDSESTSESDSESHS DSESDSDSESTSESDSESHSDSESDSDSESTSESDSESHSDSESHSDSESTSES DSESHSDSESDSDSESTSESDSESHSDSESHSDSESTSESDSESHSDSESDSDS

	ESTSESDSESHSDSESDSDSESTSESDSESHSDSESDSDSESTSESGSESHSNSE
Pro-rich LCRs	
Aap [a] (*S. epi*)	PTKAEPGKPAEPGKPAEPGKPAEPGTPAEPGKPAEPGTPAEPGKPAEPGKPAEPGKPAEPGKPAEPGTPAEPGTPAEPGKPAEPGTPAEPGKPAEPGTPAEPGKPAESGKPVEPGTPAQSGAPEQPNRSMHSTDNKNQ
SasG	PKDPKGPENPEKPSRPTHPSGPVNPNNPGLSKDRAKPNGPVHSMDKNDKVKKSKIAKESVANQEKKRAE
CNA	PEKPNKPIYPEKPKDKTPPNKPDHSNKVRPTPPDEPSKVDKVDQPKDNKTKPENPLKE
FnbpA	PPIVPPTPPTPEVPSEPETPTPPTPEVPSEPETPTPPTPEVPSEPETPTPPTPEVPAEPGKPVPPAKEEPKKPSKPVEQGKVVTPVIEINEKVKAVAPTKKPQSKKSE
FnbpB	PPIVPPTPPTPEVPSEPETPTPPTPEVPSEPETPTPPTPEVPTEPGKPIPPAKEEPKKPSKPVEQGKVVTPVIEINEKVKAVVPTKKAQSKKSE
Other LCRs	
SraP (SasA)	MSGSQSISDSTSTSMSGSTSTSESNSMHPSDSMSMHHTHSTSTSRLSSEATTSTSESQSTLSATSEVTKHNGTPAQSEKR
FmtB (SasB)	NNKATQNDGANASPATVSNGSNSANQDMLNVTNTDDHQAKTKSAQQGKVNKAKQQAKT
SasC	DTAIGQIDQDRSNAQVDKTASLNLQTIHDLDVHPIKKPDAEKTINDDLARVTALVQNYRKVSDRNKADALKAITALKLQMDEELKTARTNADVDAVLKRFNVALSDIEAVITEKENSLLRIDNIAQQTYAKFKAIATPEQLAKVKVLIDQYVADGNRMIDEDATLNDIKQHTQFIVDEILAIKLPAEATKVSPKEIQPAPKVCTPIKKEETHESRKVEKE

[a] The Aap sequence listed here is based on the consensus identification of the LCR region by the PlaToLoCo server [9], as for all other sequences in Table 5. This sequence differs slightly from the Aap construct used for experimental approaches (compare to Figure 1).

References

1. Holehouse AS, Das RK, Ahad JN, Richardson MO, Pappu RV. CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins. *Biophys J* **112**, 16-21 (2017).

2. Das RK, Pappu RV. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc Natl Acad Sci U S A* **110**, 13392-13397 (2013).

3. Sawle L, Ghosh K. A theoretical method to compute sequence dependent configurational properties in charged polymers and proteins. *The Journal of chemical physics* **143**, 085101 (2015).

4. Martin EW, Holehouse AS, Grace CR, Hughes A, Pappu RV, Mittag T. Sequence Determinants of the Conformational Properties of an Intrinsically Disordered Protein Prior to and upon Multisite Phosphorylation. *J Am Chem Soc* **138**, 15323-15335 (2016).

5. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157**, 105-132 (1982).

6. Tomasso ME, Tarver MJ, Devarajan D, Whitten ST. Hydrodynamic Radii of Intrinsically Disordered Proteins Determined from Experimental Polyproline II Propensities. *PLoS computational biology* **12**, e1004686 (2016).

7. English LR, Tilton EC, Ricard BJ, Whitten ST. Intrinsic alpha helix propensities compact hydrodynamic radii in intrinsically disordered proteins. *Proteins* **85**, 296-311 (2017).

8. Yarawsky AE, English LR, Whitten ST, Herr AB. The Proline/Glycine-Rich Region of the Biofilm Adhesion Protein Aap Forms an Extended Stalk that Resists Compaction. *J Mol Biol* **429**, 261-279 (2017).

9. Jarnot P*, et al.* PlaToLoCo: the first web meta-server for visualization and annotation of low complexity regions in proteins. *Nucleic Acids Res* **48**, W77-W84 (2020).