Supporting information for:

Identifying block shift and misplacement errors in multiple sequence alignments

Aparna Roy, Bruck Taddese, Shabana Vohra, Phani K Thimmaraju,

Christopher J.R. Illingworth, Lisa M Simpson, Keya Mukherjee,

Christopher A Reynolds* and Sree V. Chintapalli*

Department of Biological Sciences, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, United Kingdom

Supporting Figures

Fig. S1a. A subset showing the normal appearance of the KKLH motif, which in class B G protein coupled receptors takes RKLH as the preferred form. The motif residues in intracellular loop 1, hence the gap immediately after the motif is not unexpected. However, the presence of the gap can lead to the loss of a recognisable motif, as shown below.

Fig. S1b. A subset from the class B GPCR alignment. For this subset, the consensus of the KKLH motif is EKLR. The alignment contains 3 blocks shift errors as the YVA / YIA should align with the FLA / YIM to align VALR or IALR with the EKLR incarnation of the KKLH motif.

Fig. S1c. A subset from the class B GPCR alignment. For this subset, the various expressions of the KKLH motif are more apparent when the reader is aware of the misplacement of a residue within a gap, e.g. R in e1bj11_bovn and block-shift errors, e.g. CKA in emr3_human.

Fig. S1d. A subset from the class B GPCR alignment. For this subset, the various expressions of the KKLH motif, SEMR, SDMR, SGMR and SQMR are more apparent when the reader is aware of the misplacement of a residue within a gap, e.g. S (and D) in b4l8z1_dromo.

Fig. S1e. A subset from the class B GPCR alignment. For this subset, the various expressions of the KKLH motif, ESLS, PVLR, LRLN, LFLR, RNLRL, ASLK, GVLK and FLNV are more apparent when the reader is aware of the misplacement of a residue within a gap, e.g. L in b3sc56_triad and LF a7sas3_nemve.

Fig. S1f. A subset from the class B GPCR alignment. For this subset, the various expressions of the KKLH motif, e.g. RSLR are more apparent when the reader is aware of the misplacement of a residue within a gap, e.g. RS in XP_395896 and R in b4jx4_drogr. The error is possibly caused by an alignment error in the final sequence, b4gtz3_drope.

Fig S2a. A subset from the calcitonin alignment (calcitonin is a class B GPCR). The CPDH in a7z0c1_oryla is displaced 2 positions to the right. The structure is unknown as this region occurs N-terminal to any known X-ray structures for this family (terHaar E. *et al.* 2010;Kusano *et al.* 2012).

Fig S2b. A subset from the calcitonin alignment. The LDDELE... is displaced 2 positions to the right. The structure is unknown as this region occurs at the interface between the N-terminus and transmembrane helix 1.

Fig. S3. A subset from the calcitonin alignment. The YDN in *q16u26_aedae* is displaced 6 positions to the left. The CW motif is a key motif in extracellular loop 2.

Fig S4a. A subset from the calcitonin alignment. The M in intracellular loop 2 of *e0ve04_pedhc* is displaced 16 positions to the left.

Fig. S4b. A subset from the calcitonin alignment. The QIVF in intracellular loop 2 of *pdf1_caee1* is displaced 1 position to the left.

Supporting Tables

Table S1: NorMD quality score for Clustal, MUSCLE, MAFFT and Kalign alignments. The highest NorMD is indicated in **bold** and the lowest NorMD is indicated in *italics*.

Codes	Sequences	Clustal	MUSCLE	MAFFT	Kalign
1ak4C	15	1.000	1.000	1.000	1.000
1agrA	202	<i>0.630</i>	0.649	0.648	0.672
1volA	525	<i>0.371</i>	<i>0.371</i>	0.399	<i>0.398</i>
1atnD	31	<i>0.493</i>	<i>0.519</i>	0.812	<i>0.572</i>
1axiA	867	0.641	<i>0.564</i>	0.665	0.884
1e3dA	30	<i>0.390</i>	0.481	0.428	0.558
1ak4A	37	<i>0.441</i>	0.622	0.605	<i>0.551</i>
1agrE	80	<i>0.257</i>	0.406	0.485	0.503
1avaC	409	0.568	<i>0.552</i>	0.746	<i>0.734</i>
1auIB	211	0.639	0.628	<i>0.620</i>	<i>0.626</i>
1hrpA	118	<i>0.965</i>	1.178	1.187	<i>1.174</i>
1jk0B	999	0.564	<i>0.531</i>	0.672	<i>0.647</i>
2pcbA	47	<i>0.508</i>	0.582	0.891	<i>0.737</i>
1auIA	177	0.593	0.662	<i>0.643</i>	<i>0.580</i>
1hrpB	25	1.443	<i>1.361</i>	1.487	<i>1.395</i>
1jk0A	999	0.572	<i>0.512</i>	0.672	<i>0.628</i>
1f88A	152	0.662	<i>0.642</i>	0.679	<i>0.658</i>
Mean		<i>0.632</i>	0.662	0.743	<i>0.725</i>
Std. Deviation		0.280	0.267	0.274	0.255

Table S2: NorMD quality score for Clustal and T-Coffee. The highest NorMD is indicated in **bold**. The norMD score for the composite 1hrpA, 1c3aB, 1agrA and 1dkfB alignments, generated as per the T-coffee alignments was 0.124, 0.001, 0.698 and 0.314 respectively.

Codes	Sequences	Clustal	T-Coffee
1volB	544	0.602	0.545
1avaA	57	0.258	0.580
1hrpA	118	0.965 ^a	1.077
1hrpB	25	1.443	1.316
1jk0A	999	0.572	0.439
1axiB	35	0.535	0.602
1c3aB	137	0.486 ^a	0.791
1e3dB	546	0.854	0.480
1atnD	31	0.493	0.569
1agrA	202	0.630 ^a	0.718
1dkfB	110	0.577 ^a	0.631
1he8A	34	0.702	0.806
1jatB	72	0.241	0.484
Mean		0.643	0.695
Std. Deviation		0.312	0.254

^a The norMD score for the composite 1hrpA, 1c3aB, 1agrA and 1dkfB alignments, generated as per the T-coffee alignments was 0.124, 0.001, 0.698 and 0.314 respectively.

References

- Kusano,S., Kukimoto-Niino,M., Hino,N., Ohsawa,N., Okuda,K., Sakamoto,K., Shirouzu,M., Shindo,T. & Yokoyama,S. (2012). Structural basis for extracellular interactions between calcitonin receptor-like receptor and receptor activity-modifying protein 2 for adrenomedullin-specific binding. *Protein Sci.*, 21(2): 199-210.
- ter Haar E., Koth,C.M., Abdul-Manan,N., Swenson,L., Coll,J.T., Lippke,J.A., Lepre,C.A., Garcia-Guzman,M. & Moore,J.M. (2010). Crystal structure of the ectodomain complex of the CGRP receptor, a class-B GPCR, reveals the site of drug antagonism. *Structure*, 18(9): 1083-1093.