Identifying subset 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

School of Biological Sciences, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK

Communicated by Ramaswamy H. Sarma

(Received 12 November 2012; final version received 22 January 2013)

Multiple sequence alignment (MSA) accuracy is important, but there is no widely accepted method of judging the accuracy that different alignment algorithms give. We present a simple approach to detecting two types of error, namely block shifts and the misplacement of residues within a gap. Given a MSA, subsets of very similar sequences are generated through the use of a redundancy filter, typically using a 70–90% sequence identity cut-off. Subsets thus produced are typically small and degenerate, and errors can be easily detected even by manual examination. The errors, albeit minor, are inevitably associated with gaps in the alignment, and so the procedure is particularly relevant to homology modelling of protein loop regions. The usefulness of the approach is illustrated in the context of the universal but little known [K/R]KLH motif that occurs in intracellular loop 1 of G protein coupled receptors (GPCR); other issues relevant to GPCR modelling are also discussed.

Keywords: redundancy; errors; multiple sequence alignments; alignment accuracy; alignment errors; homology modelling

Introduction

Multiple sequence alignments (MSAs), play a central role in bioinformatics as they lay the foundation for identifying conserved residues, identifying functional sites in proteins (Dean et al., 2001; Gouldson et al., 2001; Lichtarge, Bourne, & Cohen, 1996; Madabushi et al., 2004; Mihailek, Res, & Lichtarge, 2004), predicting secondary structure (Cuff, Clamp, Siddiqui, Finlay, & Barton, 1998; Jones, 1999; McGuffin, Bryson, & Jones, 2000; Rost, 1996), and implying phylogeny (Felsenstein, 1989); the accuracy of the alignments is important, as different methods can lead to different conclusions in phylogenetic analysis and different structures in homology models (Dickson, Wahl, Fernandes, & Gloor, 2010; Loytynoja & Goldman, 2008; Martin, MacArthur, & Thornton, 1997; Venclovas 2001, 2003; Venclovas & Margelevicius, 2005; Venclovas, Zemla, Fidelis, & Moul, 2001; Wong, Suchard, & Huelsenbeck, 2008). The issue of accurately assessing the quality of a MSA is not straightforward (Aniba, Poch, Marchler-Bauer, & Thompson, 2011; Aniba et al., 2010b; Edgar, 2004; Raghava, Searle, Audley, Barber, & Barton, 2003). There are systematic approaches for identifying badly aligned residues in MSAs (where there is no reference to e.g. benchmarks) (Blouin, Perry, Lavell, Susko, & Roger, 2009; Dickson et al., 2010; Lassmann & Sonnhammer, 2007), and there are several approaches for scoring alignments and for removing badly aligned sequences or correcting poorly aligned regions (Lassmann & Sonnhammer, 2005a; Muller, Creevey, 2010).

In principle, dynamic programming (Needleman & Wunsch, 1970) permits the determination of an optimum alignment, but for more than about 10–20 sequences, CPU demands dictate the use of heuristic approaches, as in Clustal (Larkin et al., 2007; Thompson, Gibson, Plewniak, Jeanmougin, & Higgins, 1997). However, the optimum defined by the objective function does not necessarily guarantee that the MSA has optimal biological meaning, making it difficult to assess MSAs. In this multiple sequence context, reliability is usually sought by choosing an alignment program that performs well against a series of curated MSA benchmarks such as BAliBASE, Oxbench and Prefab (Aniba, Poch, Marchler-Bauer, & Thompson, 2011a; Aniba et al., 2010b; Edgar, 2004; Raghava, Searle, Audley, Barber, & Barton, 2003). There are systematic approaches for identifying badly aligned residues in MSAs (where there is no reference to e.g. benchmarks) (Blouin, Perry, Lavell, Susko, & Roger, 2009; Dickson et al., 2010; Lassmann & Sonnhammer, 2007), and there are several approaches for scoring alignments and for removing badly aligned sequences or correcting poorly aligned regions (Lassmann & Sonnhammer, 2005a; Muller, Creevey, 2010).
Thompson, Arendt, & Bork, 2010; Thompson, Thierry, & Poch, 2003; Thompson, Plewniak, Ripp, Thierry, & Poch, 2001). The approach presented here, inspired by the work of Baldwin (Baldwin, 1993; Baldwin, Schertler, & Unger, 1997), is fundamentally different as the focus is on generally well-aligned subsets of very similar sequences and while the errors are relatively minor, they are nevertheless distinctly unwanted in some applications such as homology modelling. It involves splitting the MSA into subsets of sequences with high sequence identity, in which subsets errors can be easily identified. Here we have analysed alignments from a number of sources to illustrate the type of errors and the error rate that may be expected in both database alignments and in alignments generated in house.

Methods

Multiple sequence alignments

MSAs were generated with default parameters for 17 protein sequences of known structure, collected following a BLAST search (Altschul et al., 1997) terminated at an E value of 0.001 using ClustalX v1.8 (Thompson, Higgins, & Gibson, 1994; Thompson et al., 1997), Kalign v2.04 (Lassmann & Sonnhammer, 2005b), MaFFT (Katoh, Misawa, Kuma, & Miyata, 2002), Muscle v3.7 (Edgar, 2004) and T-coffee v7.97 (Notredame, Higgins, & Heringa, 2000). Our version of T-coffee did not process more than 100 sequences and so for T-coffee MSAs of greater than 100 sequences, a Clustal alignment was initially used and the sequences were then ordered according to sequence identity using Jalview (Clamp, Cuff, Searle, & Barton, 2004). The alignment was then split into groups of 99 sequences that were aligned using T-coffee and these alignments were then merged using a profile alignment in Clustal (in order of decreasing sequence identity). In some cases, alignments were improved using RASCAL (Thompson et al., 2003) and scored using norMD (Thompson et al., 2001). The quality of the MSAs was evaluated using norMD, but it should be stressed that norMD is assessing different aspects of the MSA quality to those described below. In addition, a MSA of 1000 class B G protein coupled receptors (GPCRs) was obtained from the GPCRDB (Horn et al., 1998) and realigned using Clustal; 50 subsets were generated at the 70% sequence identity level. An alignment of 278 sequences from the class B calcitonin family was generated and analysed in a similar way.

Generating subsets to identify subset errors

The MSAs were processed through the ExPASy ‘remove redundancy’ filter (Notredame, 2007), or in-house code, to remove sequences with 90% and 70% sequence identity. Together the sequences removed, along with the sequence left in the alignment to which they were similar, formed subsets of similar (or very similar in the case of 90% identity) sequences. Because of the high similarity within the subsets, the errors in the alignment were identified very readily. Such subset errors are clearly visible in Figure 1(a). While we show here that these errors can be significant, they usually appear in less important loop regions and have previously largely

Figure 1. A subset from the class B GPCR alignment. (a) For this subset, the various expressions of the [K/R]KLH motif are RSLK, RTR, RPIQ, RMLR, RALR, RAMK, RTLR, SNLH, RTLR, RGLK, HKLR, RALR, KGLK, RALR and RMLR. The motif is more apparent when the reader is aware of the misplacement of a residue within a gap, for example, R in XP_001494282 and block-shift errors, for example, in b7z7u7_human. (b) Errors in the subset have been corrected using Jalview to reveal errors in sequences a4kp32_danre and a9c3t8_danre. (c) The gap in the helical region has been removed by realigning sequences a4kp32_danre and a9c3t8_danre.
been ignored as they can appear trivial, particularly as they do not usually affect the NORMD score.

Results and discussion

Typical errors

Typical block shift and misplacement errors in the [R/K] KLH motif are shown in Figure 1(a). The errors involve the mis-placement of a residue, which commonly occurred either side of a gap, for example, R in XP-001,494,282, q4rxp0_dadre, cchl2_rat and XP002751246, which should move to the right hand side of the gap to complete the motif (and thus align with other Rs in the same column). The errors also involve shifts in a short block of residues e.g. RML in cchl1_human and RTL in q57119_mouse (so that the R and L can align on the right hand side). In this subset, shifts of short blocks of 2 or 4 residues are also observed, for example, RS in q5y190_human and RM in v7z7u7_human, RTLR in XP_001196967 and XP_002661517 and KHLR. Sequence q4kp32_danre has two errors. Frame shift errors (block shift errors that continue to the end of the sequence) can also be observed but these are far less common. In general, such alternative positions, for example, for R in Figure 1(a), would be equally viable, especially if both columns contain R. When analysed within the context of a MSA, these errors are largely invisible. However, they become more apparent when subsets are analysed. Thus, Figure S1 shows a subset where the motif aligns perfectly, along with other subsets that contain similar errors to those in Figure 1(a). However, when the errors occur in highly similar subsets it is imperative that the alignment is corrected, because of the very nature of degenerate sequences.

We note that a block shift of 1 residue could give rise to an error of over 5 Å in the position of a Cβ atom in a homology model. Since the errors are inevitably associated with gaps, this procedure of error detection is particularly relevant to homology modelling when the loop regions of the modelled protein are of interest, since gaps are frequently more prominent in loop regions. This homology modelling problem is well illustrated with the class B GPCR alignments, which contain a [K/R] KL1.63H motif in intracellular loop 1 (Figures 1 and S1); this motif is universal (Vohra et al., 2013) in the sense that it is also seen in class E and class A GPCRs, where the X-ray structures containing this motif all have a well-defined structure in which the L (or similar) at position 1.63 (generic numbering (Ballesteros & Weinstein, 1995)) makes an interaction with F (or similar) at position 8.50 on helix 8, which also contacts Y7.53 (or similar) in the inactive structures; this hydrophobic interaction is often accompanied by a polar interaction between 1.61 or 1.62 and 8.49 (Vohra et al., 2013). Although the motif is not well-known within class A or class B GPCRs as the percentage identity is not high, it came to light in an alignment of class A and class B GPCRs where it provides one of the more obvious similarities between the classes (Vohra et al., 2013). Its role in GPCR structure and stability will be discussed elsewhere (Simpson et al., 2013). In many of the 50 subsets, the motif is well aligned, and various alternative forms of the motif are apparent, for example, RRLH, RSLR, SNLH (Figure S1, Table 1). In some subsets, the motif is split by gaps (with no block shift errors etc., as in Figure S1a); this may be one of the reasons why this motif has not previously been widely discussed in the literature. In a small proportion of these subsets, however, there are block shift errors, as discussed above and shown in Figure 1(a) and Figures S1b-S1f, and this could disrupt a homology model. The most likely outcome in homology modelling is that the very specific structural form of the KKLH motif, typically seen in rhodopsin, is replaced by an unstructured loop.

The [R/K]KLH motif is not the only place where these errors occur. Additional errors in the alignment of the N-terminus of the calcitonin receptor are shown in Figure S2. It is difficult to say anything about the error in Figure S2a as it lies in a region where there is no structural information as the sequences of the known structures within this family are not as long as the ones in Figure S2a. The error in Figure S2b lies at the interface between the N-terminus and transmembrane helix 1. There is interest in modelling class B GPCR structures by combining models of the transmembrane region with X-ray crystal structures of the N-terminus (Dong et al., 2011; Parthier, Reedtz-Runge, Rudolph, & Stubbs, 2009). However, modelling the hinge region of these models could be limited by errors such as those in Figure S2b.

An error in the alignment of extracellular loop 2 (EL2) is shown in Figure S3. EL2 plays a key role in receptor activation and ligand binding (Barwell et al., 2013; Ivanov, Barak, & Jacobson, 2009) and errors in EL2 have been shown to limit the development and use of GPCR homology models (Ivanov et al., 2009), which can otherwise be useful in both the study of receptor activation and in docking studies to identify novel ligands (Carlsson et al., 2010; Katrich et al., 2010; Taddei, Simpson, Wall, Blaney, & Reynolds, 2013).

Errors that could result in insertions in intracellular loop 2/transmembrane helix 4 are shown in Figure S4; incorrect treatment of such errors may leave the helix or adjoining loop out of register, with key residues facing in the wrong direction. While deletions in helix 2 of class A GPCRs have been proposed as major evolutionary events (Chabbert et al., 2012), similar subsets should not differ in this regard. The GWG motif (Figure S4a) or its GYG equivalent (Figure S4b) are well known helical motifs (Vohra et al., 2013; Lagerstrom & Schioth 2008).
While the best place to insert the gap in Figures S4a and S4b. may not be obvious, it is probably better to insert the gap sufficiently to the left of the GWG/GYG motif to ensure that it is in the loop region. However, this
region of the receptor couples to the G protein (Rasmussen et al., 2011) so the gap should be inserted with care.

Comparing alignment programs

The ranking of the alignment programs, as judged by the total number of subsets with errors, is Kalign > MAFFT Clustal > Muscle (Table 2). The order is surprising as Kalign, Muscle and MAFFT were developed to align large numbers of sequences rather than with accuracy as the sole criterion. The order presented here is not the same as that presented elsewhere (Katoh et al., 2002; Lassmann & Sonnhammer, 2005b), because the criteria are very different: here, we are assessing only the small number of subset errors that occur in very similar sequences rather than an alignment score based on the whole MSA that may be heavily influenced by divergent sequences. In general, the most appropriate alignment algorithm depends on the nature of the sequences (Thompson, Plewniak, & Poch, 1999) and the choice of alignment program should not be made on the basis of these results, but the results should influence the effort made to identify and correct subset errors. The norMD ranking (MAFFT > Kalign > Muscle, Table S1) does not necessarily correlate with the number of subset errors because norMD scores the complete MSA whereas we only assess a small part of the MSA. The widely used Clustal program generally had a low error rate, but Kalign is unique in having almost no block shift or misplacement errors. It is interesting to note that different numbers of subsets and norMD scores were generated for each alignment even though the slightly different alignments were all generated from the same sequences.

The comparison between Clustal and T-coffee, given in Table 2, was not limited to alignments of 50/100 sequences but rather larger T-coffee alignments were determined as described in the methods section. Here, T-coffee generally performs considerably better than Clustal according to the norMD scores in Table S2, but Clustal gave fewer subset errors than T-coffee. Surprisingly, one of the Clustal alignments (1agrA) constructed by combining shorter Clustal alignments using profile alignments (as for T-coffee), gave a higher norMD score (0.698, Table S2) than that obtained for the standard Clustal alignment (0.630) – possibly because the propagation of errors throughout the alignment was limited.

Errors are more likely to be observed when the sequence identity level is decreased, for example, from 90 to 70%, as shown in Table 3. Even more errors are likely to be observed when the sequence identity level is decreased, for example, down to 50 or 40%, (results not shown), but under these circumstances the method is far less valid as the sequences are no longer highly similar and structural alignments and sequence alignments are unlikely to be equivalent. However, we note that a large block shift might cause a sequence to be incorrectly omitted from a given subset, whereas it might be identified when the sequence identity criterion is relaxed – for

| Codes   | Sequs | Clustal (90%) |   | Muscle (90%) |   | MAFFT (90%) |   | Kalign (90%) |   |
|---------|-------|--------------|---|--------------|---|-------------|---|--------------|---|
|         |       | Subsets Errors |   | Subsets Errors |   | Subsets Errors |   | Subsets Errors |   |
| 1ak4C   | 15    | 1 0          |   | 1 0           |   | 1 0          |   | 1 0          |   |
| 1agrA   | 202   | 33 0         |   | 33 4          |   | 33 0         |   | 35 0         |   |
| 1volA   | 525   | 82 0         |   | 83 1          |   | 90 0         |   | 83 0         |   |
| 1tnID   | 31    | 5 5          |   | 5 0           |   | 5 0          |   | 5 0          |   |
| 1axIA   | 867   | 86 11        |   | 105 22(28)    |   | 100 10(15)   |   | 94 0         |   |
| 1e3DA   | 30    | 4 0          |   | 5 0           |   | 5 0          |   | 4 0          |   |
| 1ak4A   | 37    | 7 0          |   | 7 0           |   | 7 0          |   | 7 0          |   |
| 1agRE   | 80    | 24 0         |   | 17 1          |   | 16 0         |   | 18 0         |   |
| 1avaC   | 409   | 54 1         |   | 50 23(44)     |   | 53 11(17)    |   | 50 1         |   |
| 1auIB   | 211   | 34 1         |   | 27 3          |   | 32 1         |   | 31 0         |   |
| 1hrpA   | 118   | 14 0         |   | 14 1          |   | 15 0         |   | 14 0         |   |
| 1jk0B   | 999   | 132 9        |   | 132 59(210)   |   | 131 14(16)   |   | 136 0        |   |
| 2pcBA   | 47    | 5 1          |   | 5 0           |   | 6 0          |   | 5 0          |   |
| 1auIA   | 177   | 33 3(5)      |   | 26 9(12)      |   | 32 2         |   | 30 0         |   |
| 1hrpB   | 25    | 1 0          |   | 1 0           |   | 2 0          |   | 1 0          |   |
| 1jk0A   | 999   | 139 19       |   | 138 84(318)   |   | 129 19(23)   |   | 134 0        |   |
| 1f88A   | 152   | 22 0         |   | 19 3          |   | 19 0         |   | 24 0         |   |
| Total   | 43    |              |   | 210           |   | 57           |   | 1            |   |
this reason a poor alignment could score better than a good alignment by our criteria.

Correcting MSAs

A simple way to correct the errors would be to edit the subsets and add them back to the whole MSA using a profile alignment. Thus, Figure 1(b) shows corrections to the alignment shown in Figure 1(a) enacted by using Jalview as an editor by (i) moving the various parts of the KKLH motif to coincide with the correct form of the motif as shown in Figure S1a and (ii) removing a single gap from each sequence between helix 1 and the motif (~position 10). This leaves a gap of three residues in two sequences (a4kp32 danre and a9c3t8 danre) at around position 10, which is problematic because it occurs in the helix. This second error is only apparent when the first error is corrected; the second error can be fixed by moving the helical residues three places to the right, as shown in Figure 1(c). This subset can now be added back to main alignment using a profile alignment.

Ideally, subsets should be generated at multiple degeneracy levels to identify all the block-shift and misplacement errors. For example, in the 1he8A alignment, the Clustal alignment contains one unique error that can be detected at any degeneracy level between 40 and 85%, while the T-coffee alignment contains three unique errors, one of which can be detected at 65–85%, 2 can be detected at 50–60%, while 3 can be detected at 40–45%. Different alignment programs insert gaps at different places – hence the different number of errors for the different methods in Tables 2 and 3. Corrections to errors where a residue is misplaced within a gap should be undertaken with reference to similar residues within both the subset and the whole MSA but they should clearly be in the same position within a highly similar subset.

Conclusions

Here we have presented a simple procedure for identifying block shift and misplacement errors in MSAs by the analysis of subsets of (highly) similar sequences in which the determination of the errors is relatively straightforward. The approach has been applied to a number of MSAs generated using Clustal, Kalign, MaFFT, Muscle and T-coffee. MSAs generated using Kalign had almost no block shift and misplacement errors while those generated using Muscle and T-coffee had the most. However, these errors are not the sole determinant of alignment quality as alignments with higher norMD scores sometimes had more block shift and misplacement errors. For this reason, the method is complementary to those that assess the whole alignment, including the more divergent sequences. The block shift and misplacement errors are inevitably associated with gaps and so may pose a threat to the loop regions of homology models since such loop regions generally correspond to regions of a MSA that are rich in gaps. It is in homology modeling where minor alignment errors can have large deleterious effects that this method is likely to be most beneficial. Moreover, it is possible that the [R/K]KLH motif in intracellular loop 1 of class A class B and class R GPCRs, has not been previously discussed because these errors have helped to obscure it. The procedure for correcting these errors will depend on the degree of degeneracy within the subset. For highly similar subsets, identical residues involved in misplacement errors should almost certainly reside at the same position.
position; the alignment should be edited with reference to similar residues within both the subset and the whole MSA.

Supplementary material
The supplementary material for this paper is available online at http://dx.doi.org/10.1080/07391102.2013.770371.

References
Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 25, 3389–3402.

Aniba, M. R., Poch, O., Marchler-Bauer, A., & Thompson, J. D. (2010a). AlexSys: A knowledge-based expert system for multiple sequence alignment construction and analysis. *Nucleic Acids Research*, 38, 6338–6349.

Aniba, M. R., Poch, O., Marchler-Bauer, A., & Thompson, J. D. (2010b). AlexSys: A knowledge-based expert system for multiple sequence alignment construction and analysis. *Nucleic Acids Research*, 38, 6338–6349.

Baldwin, J. M. (1993). The probable arrangement of the helices in G-protein-coupled receptors. *EMBO Journal*, 12, 1693–1703.

Baldwin, J. M., Schertler, G. F., & Ungr, V. M. (1997). An alpha-carbon template for the transmembrane helices in the rhodopsin family of G-protein-coupled receptors. *Journal of Molecular Biology*, 272, 144–164.

Ballesteros, J. A., & Weinstein, H. (1995). Integrated methods for the construction of three-dimensional models and computational probing of structure-function relationships in G-protein coupled receptors. *Methods Neurosci.*, 25, 366–428.

Barwell, J., Wheatley, M., Conner, A. C., Taddese, B., Vohra, S., Reynolds, C. A., & Poyner, D. R. (2013). The activation of the CGRP receptor. *Biochemistry Society Transactions*, 41, 180–184.

Blouin, C., Perry, S., Lavel, A., Susko, E., & Roger, A. J. (2009). Reproducing the manual annotation of multiple sequence alignments using a SVM classifier. *Bioinformatics*, 25, 3093–3098.

Carlsson, J., Yoo, L., Gao, Z. G., Irwin, J. J., Shoichet, B. K., & Jacobson, K. A. (2010). Structure-based discovery of A2A adenosine receptor ligands. *Journal of Medicinal Chemistry*, 53, 3748–3755.

Chabbert, M., Castel, H., Pele, J., Deville, J., Legendre, R., & Rodien, P. (2012). Evolution of class a G-protein-coupled receptors: Implications for molecular modelling. *Current Medicinal Chemistry*, 19, 1110–1118.

Clamp, M., Cuff, J., Searle, S. M., & Barton, G. J. (2004). The Jalview Java alignment editor. *Bioinformatics*, 20, 426–427.

Cuff, J. A., Clamp, M. E., Siddiqui, A. S., Finlay, M., & Barton, G. J. (1998). JPred: A consensus secondary structure prediction server. *Bioinformatics*, 14, 892–893.

Dean, M. K., Higgs, C., Smith, R. E., Bywater, R. P., Snell, C. R., Scott, P. D., … Reynolds, C. A. (2001). Dimerization of G-protein-coupled receptors. *Journal of Medicinal Chemistry*, 44, 4595–4614.

Dickson, R. J., Wahl, L. M., Fernandes, A. D., & Gloor, G. B. (2010). Identifying and seeing beyond multiple sequence alignment errors using intra-molecular protein cavitation. *PLoS One*, 5, e11082.
Martin, A. C., MacArthur, M. W., & Thornton, J. M. (1997). Assessment of comparative modeling in CASP2. *Proteins, 1*, 14–28.

McGuffin, L. J., Bryson, K., & Jones, D. T. (2000). The PSI-PRED protein structure prediction server. *Bioinformatics, 16*, 404–405.

Mihalek, I., Res, I., & Lichtarge, O. (2004). A family of evolution-entropy hybrid methods for ranking protein residues by importance. *Journal of Molecular Biology, 336*, 1265–1282.

Muller, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology, 48*, 443–453.

Needleman, S. B., & Wunsch, C. D. (1970). A general method for fast and accurate multiple sequence alignment. *Bioinformatics, 4*, 263–266.

Notredame, C. (2007). Decrease redundancy server at www.expasy.org/tools (unpublished).

Notredame, C., Higgins, D. G., & Heringa, J. (2000). T-Coffee: a novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology, 302*, 205–217.

Parthier, C., Reedtz-Runge, S., Rudolph, R., & Stubbs, M. T. (2009). Passing the baton in class B GPCRs: peptide hormone activation via helix induction? *Trends in Biochemical Sciences, 34*, 303–310.

Raghava, G. P., Searle, S. M., Audley, P. C., Barber, J. D., & Barton, G. J. (2003). OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. *BMC Bioinformatics, 4*, 47.

Rasmussen, S. G., Devree, B. T., Zou, Y., Kruse, A. C., Chung, K. Y., Kobilka, T. S., … Kobilka, B. K. (2011). Crystal structure of the beta2 adrenergic receptor-Gs protein complex. *Nature, 477*, 549–555.

Rost, B. (1996). PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods In Enzymology, 266*, 525–539.

Simpson, L. M., Wall, I. D., & Reynolds, C. A. (submitted for publication). A computational approach to assessing GPCR stability. *MedChemComm.*

Taddese, B.; Simpson, L. M., Wall, I. D., Blaney, F. E., & Reynolds, C. A. (2013). Modeling GPCR active state conformations. *Methods in Enzymology, 522*, 21–35.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., & Higgins, D. G. (1997). The CLUSTAL X windows interface. flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research, 25*, 4876–4882.

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research, 22*, 4673–4680.

Thompson, J. D., Plewniak, F., & Poch, O. (1999). A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Research, 27*, 2682–2690.

Thompson, J. D., Plewniak, F., Ripp, R., Thiery, J. C., & Poch, O. (2001). Towards a reliable objective function for multiple sequence alignments. *Journal of Molecular Biology, 314*, 937–951.

Thompson, J. D., Thiery, J. C., & Poch, O. (2003). RASCAL: rapid scanning and correction of multiple sequence alignments. *Bioinformatics, 19*, 1155–1161.

Venclovas, C. (2001). Comparative modeling of CASP4 target proteins: combining results of sequence search with three-dimensional structure assessment. *Proteins, 5*, 47–54.

Venclovas, C. (2003). Comparative modeling in CASP5: progress is evident, but alignment errors remain a significant hindrance. *Proteins, 53*, 380–388.

Venclovas, C., & Margelevicius, M. (2005). Comparative modeling in CASP6 using consensus approach to template selection, sequence-structure alignment, and structure assessment. *Proteins, 61*, 99–105.

Venclovas, C., Zemla, A., Fidelis, K., & Moult, J. (2001). Comparison of performance in successive CASP experiments. *Proteins, 5*, 163–170.

Vohra, S., Taddese, B., Conner, A. C., Poyner, D. R., Hay, D. L., Barwell, J., … Reynolds, C. A. (2013). Similarity between class A and class B G-protein-coupled receptors exemplified through calcitonin gene-related peptide receptor modelling and mutagenesis studies. *Journal of the Royal Society Interface, 10*, 20120846.