Re-evaluation of low-resolution crystal structures *via* interactive molecular-dynamics flexible fitting (iMDFF): a case study in complement C4

Tristan Ian Croll^{a*} and Gregers Rom Andersen^b

Received 10 May 2016

Accepted 27 July 2016

Edited by K. Diederichs, University of Konstanz, Germany

Keywords: interactive molecular dynamics flexible fitting; model building; stereochemical quality; low-resolution structures; complement C4.

PDB references: complement C4, 5jpn; complex of complement C4 and MASP-2, 5jpm; complement C4b, 5jtw

Supporting information: this article has supporting information at journals.iucr.org/d

^aInstitute of Health and Biomedical Innovation, Queensland University of Technology, GPO Box 2434, Brisbane, QLD 4001, Australia, and ^bDepartment of Molecular Biology and Genetics, Aarhus University, Gustav Wieds Vej 10C, 8000 Aarhus, Denmark. *Correspondence e-mail: tristan.croll@qut.edu.au

While the rapid proliferation of high-resolution structures in the Protein Data Bank provides a rich set of templates for starting models, it remains the case that a great many structures both past and present are built at least in part by hand-threading through low-resolution and/or weak electron density. With current model-building tools this task can be challenging, and the *de facto* standard for acceptable error rates (in the form of atomic clashes and unfavourable backbone and side-chain conformations) in structures based on data with $d_{\max}$ not exceeding 3.5 Å reflects this. When combined with other factors such as model bias, these residual errors can conspire to make more serious errors in the protein fold difficult or impossible to detect. The three recently published 3.6–4.2 Å resolution structures of complement C4 (PDB entries 4fxg, 4fxk and 4xam) rank in the top quartile of structures of comparable resolution both in terms of R_{free} and *MolProbity* score, yet, as shown here, contain register errors in six β -strands. By applying a molecular-dynamics force field that explicitly models interatomic forces and hence excludes most physically impossible conformations, the recently developed interactive molecular-dynamics flexible fitting (iMDFF) approach significantly reduces the complexity of the conformational space to be searched during manual rebuilding. This substantially improves the rate of detection and correction of register errors, and allows user-guided model building in maps with a resolution lower than 3.5 Å to converge to solutions with a stereochemical quality comparable to atomic resolution structures. Here, iMDFF has been used to individually correct and re-refine these three structures to *MolProbity* scores of <1.7, and strategies for working with such challenging data sets are suggested. Notably, the improved model allowed the resolution for complement C4b to be extended from 4.2 to 3.5 Å as demonstrated by paired refinement.

1. Introduction

1.1. Biology of complement C4

The complement system is an essential arm within innate immunity, providing us with immediate defence against pathogens. Additionally, it maintains tissue homeostasis by the continuous removal of dying cells, stimulates adaptive immunity and plays key roles in development (Bajic *et al.*, 2015). Complement is based on a proteolytic cascade that is initiated upon pattern recognition of pathogen/danger-associated molecular patterns presented on the surface of pathogens, dying cells, damaged tissue and immune complexes. Surfaces presenting these patterns are collectively referred to as activators. The lectin and the classical pathways are two homologous complement initiation pathways within the

complement system. In the classical pathway (CP), IgG or IgM in immune complexes and other ligands trigger activation (Kojouharova *et al.*, 2010), whereas activation of the lectin pathway (Lp) occurs upon pattern recognition of glycan presented by pathogens or altered self-tissue (Kjaer *et al.*, 2013). Pattern recognition activates one of two proteolytic enzymes (C1s in the CP and MASP-2 in the Lp) leading to cleavage of complement C4 (hereafter referred to as C4) into C4b (195 kDa) and C4a (9 kDa). Whereas the small fragment has no established biological function, C4b is an opsonin and attaches to surface nucleophiles on the activator through their reaction with a thioester exposed upon C4 cleavage (Law, 1983). The zymogen C2 may then bind C4b to form the proconvertase C4b–C2, in which C2 becomes activated through proteolytic cleavage by C1s, MASP-2 or MASP-1, resulting in the formation of the active C3 convertase C4b–C2a. Strong amplification of the initial C4 cleavage event now occurs through C3 cleavage in the downstream alternative pathway (Bajic *et al.*, 2015).

The C4 precursor encompasses 1744 residues, and undergoes a complex maturation including the generation of the three chains α , β and γ , the addition of four N- and one O-linked glycans, and the sulfation of three tyrosine residues (Halim *et al.*, 2013; Hortin *et al.*, 1986; Fig. 1b). However, the most important post-translational modification is the formation of the internal thioester between the side chains of Cys1010 and Gln1013 (Law *et al.*, 1980). To complicate matters further, two C4 isotypes exist derived from the genes *C4A* and *C4B*, differing in only four residues (C4A, ₁₁₂₀PCPVLD₁₁₂₅; C4B, ₁₁₂₀LSPVIH₁₁₂₅), and the C4A isotype Cys1121 may become cysteinylated (Mortensen *et al.*, 2015). Owing to the presence of a variable number of *C4A* and *C4B* gene copies, the absolute level and the ratio of two isoforms in humans is rather variable. On top of the two isotypes, sequencing has revealed a total of 23 single-residue polymorphisms in human C4A and C4B (Blanchong *et al.*, 2001).

We have recently described the structures of C4 and C4 bound to a fragment of the serine protease MASP-2 (Kidmose *et al.*, 2012). C4 is organized into 12 structural domains, with eight macroglobulin (MG) domains forming the core of the molecule. The MG domains typically encompass ~100 residues organized into two antiparallel β -sheets. The thioester (TE) domain is a large α -helical domain harbouring the internal thioester. It is inserted into the CUB (complement C1r/C1s, Uegf, Bmp1) domain consisting of two antiparallel β -sheets and located between the MG7 and MG8 domains. The small C4a domain is a four-helical bundle wedged between the MG3, MG6 and MG8 domains, and finally the C-terminal C345c domain adopting a netrin fold is located on top of MG7 and MG8. We also determined the crystal structure of C4b, in which C4a has been released from C4 upon MASP-2 or C1s cleavage of C4 (Mortensen *et al.*, 2015). This structure revealed a major rearrangement, relocating in particular the MG7, CUB and the TE domains and resulting in the exposure of the thioester, making it accessible for nucleophilic attack.

1.2. Challenges in low-resolution model building

It is very well established that as the resolution of a crystallographic data set degrades, the challenge involved in fitting a reasonable model to the electron density increases rapidly (Headd *et al.*, 2012; DiMaio *et al.*, 2013; LuCore *et al.*, 2015). This may be viewed mathematically as a reduction in the observation-to-parameter ratio, and the practical outcome of this is that decreasing resolution necessitates the application of more and more prior knowledge, *e.g.* stereochemical restraint libraries and/or reference to higher resolution structures of the same or similar proteins.

In the absence of high-resolution reference structures, at a resolution lower than about 3.5 Å the task of finding the 'true' model that best fits the data rapidly becomes intractable to automated routines (although this limit continues to be pushed back; DiMaio *et al.*, 2013) and very difficult for even the most experienced structural biologist. At these resolutions data-driven placement of individual atoms is impossible, and there is no longer sufficient detail in the maps to definitively identify most side chains. In these cases decisions on the register of secondary-structure elements are typically made using some combination of inference to (often distantly) homologous structures through sequence and structural alignment, *ab initio* secondary-structure prediction, the chemical environment of the side chains in question and assessment of the real-space correlation of multi-residue stretches to the density. While often effective, this approach can be insufficient in various situations:

Figure 1

Box-and-whisker plots showing the distribution of common measures of error in stereochemistry amongst 3–4 Å resolution structures in the Protein Data Bank as of October 2015, extracted from their accompanying validation reports (Velankar *et al.*, 2016). Red numerals and dashed lines indicate the standard that is generally considered acceptable for atomic resolution structures.

where relatively short secondary-structure elements are flanked by poorly defined loops, where there is limited variation in side-chain size or properties, where individual side chains are flexible and ill-defined, and/or where the density is complicated by crystal contacts. Errors arising from incorrect register assignment in these cases may be obscured by the large numbers of smaller-scale residual errors that are common in structures of this resolution (Fig. 1).

The 3.8 Å resolution structure of the insulin receptor (IR) ectodomain (McKern *et al.*, 2006) was one case in which errors of this nature were recently identified. The IR contains three fibronectin type III domains with β -strands of approximately eight residues, many of which are flanked by extended loops and one of which makes crystal contacts with two co-crystallized antibody F(ab) fragments. Register errors in four β -strands affecting two of these domains were recently identified in a newer 3.3 Å resolution data set and corrected (alongside substantial improvement of the overall stereochemistry) with the aid of a new approach to model building termed interactive molecular-dynamics flexible fitting (iMDFF; Croll *et al.*, 2016). This approach makes use of the CHARMM36 molecular-dynamics force field (Huang & MacKerell, 2013) implemented in *VMD* (Humphrey *et al.*, 1996) and *NAMD* (Phillips *et al.*, 2005) to allow direct physical interaction with the structure *via* a haptic interface to guide localized molecular-dynamics (MD) simulations into improved correspondence with the map, while explicitly ensuring compatibility with the local physical environment.

The IR structure was solved in 2006 with stereochemistry that reflects the challenge of fitting a structure of this size and resolution using the tools available at the time. While considered quite standard according to contemporaneous metrics, the vastly improved validation tools available today (Chen *et al.*, 2010; Hoofst *et al.*, 1996; based upon a greatly expanded population of high-resolution exemplar structures) place it in the bottom quartile of structures of comparable resolution in terms of stereochemical quality. It is perhaps not surprising to find errors in register in such a structure, but what of structures of comparable resolution built and refined to more modern standards? Using iMDFF, exploration of the 3.6 Å resolution C4 structure described above (PDB entry 4fxk; Kidmose *et al.*, 2012; published in 2012 with an 81st percentile *MolProbity* score of 3.26) revealed two clear register errors affecting four β -strands in the CUB domain and tentatively supported a third register shift affecting a weakly resolved β -strand in the MG3 domain. Further remodelling and re-refinement of this, the 3.75 Å resolution C4–MASP-2 complex (PDB entry 4fxg; Kidmose *et al.*, 2012) and the 4.2 Å resolution C4b structure (PDB entry 4xam; Mortensen *et al.*, 2015) confirmed register shifts in these and a sixth strand, and notably supported extending the resolution cutoff for PDB entry 4xam from the published 4.2 Å to the maximum resolution of the data available, 3.5 Å. Here, we present the resulting three structures refined to *MolProbity* scores better than 1.7 and introduce further advances to the iMDFF methodology. We finally discuss general strategies for dealing with similarly challenging low-resolution structures and briefly

discuss the biological implications of the changes to the three C4 structures.

2. Methods

Model building in iMDFF was carried out essentially as described previously (Croll *et al.*, 2016), with additions to the methodology as follows. Improvements in NAMD and computational hardware allowed desktop-scale interactive simulations of up to approximately 300 residues at a time under implicit solvent conditions, a significant improvement over the *in vacuo* conditions used previously. Backbone φ and ψ angles were mapped against the Ramachandran plot (Lovell *et al.*, 2003) every five time steps and assessed as outliers ($P \leq 0.05\%$), allowed ($0.05\% < P \leq 2\%$) or preferred ($P > 2\%$). The results were applied as colours (red, white or blue respectively) to the C^α atoms to apply a real-time visual measure of local backbone quality. Secondary structure was updated every 50 time steps and applied to a cartoon representation of the backbone. The ability to provide scripted forces to the running simulation was added, allowing the interactive addition and removal of harmonic restraints on (x , y , z), interatomic distances and dihedral angles. In each case the applied force was calculated at every time step and was proportional to the square of the distance to the target, limited to a maximum value to avoid instability in the simulation. These primitive forces were further used to build routines to impose specific rotamers (Lovell *et al.*, 2000) on individual residues, force contiguous stretches into specific secondary structures or induce register shifts in α -helices or β -strands.

In a typical round of model building and refinement, a set of three real-space maps covering the asymmetric unit plus 10 Å padding [$2mF_o - DF_c$, $2mF_o - DF_c$ with sharpening_b_factor=80 ($2mF_o - DF_{c,sharp}$), $mF_o - DF_c$] is first generated using *phenix.maps*. Given that in a typical iMDFF round every atom will spend some time sampling the map, at this step it is very important to ensure that free reflections are excluded from map generation by setting `exclude_free_r_reflections=True`. Additionally, missing F_{obs} were filled in with F_{calc} by setting `fill_missing_f_obs=True`. For initial model-building purposes the standard grid resolution (`grid_resolution_factor=1/4`) is adequate, but final fine-scale adjustments appear to benefit from doubling the resolution (`grid_resolution_factor=1/8`). Inspection and rebuilding as necessary is typically accomplished in a series of interactive simulations each involving a contiguous 50–100 residues plus all surrounding residues which approach within 5 Å of this stretch. Maps are masked to within a user-defined range from the mobile atoms to aid visualization (a cutoff of 3–5 Å for $2mF_o - DF_c$ maps and 5–10 Å for the $mF_o - DF_c$ map is useful for most purposes) and applied to the simulation with individually defined coupling constants. These coupling constants are somewhat arbitrary, but values of 0.6, 0.3 and 0.1 for the $2mF_o - DF_c$, $2mF_o - DF_{c,sharp}$ and $mF_o - DF_c$ maps, respectively, appear to strike a useful balance in which atoms are attracted into the maps without introducing severely unnatural geometry. Problematic regions requiring substantial

rearrangement are best handled in more localized simulations in order to improve simulation speed and protect surrounding well ordered residues from accidental disruption. The complete structure is subjected to energy minimization with restraints on chirality and peptide-bond geometry released prior to writing coordinates for reciprocal-space refinement.

Since the CHARMM36 force field currently contains no parameters for organic lead compounds, for the purposes of iMDFF we replaced the trimethyllead with a simple Pb^{2+} ion using the parameters derived by Won (2012) (included as residue Pb2p in the most recent CHARMM36 force-field

release in `stream/misc/toppar_ions_won.str`). No attempt was made to model the methyl groups into the density. Lead ions were placed only where approximately spherical blobs of $>3\sigma$ appeared in the $mF_o - DF_c$ maps adjacent to acidic residues, with centroids consistent with the expected 2.5–3 Å interatomic distance to oxygen (Holden & Rayment, 1991). Pb–O bond distances were restrained to 2.75 Å during refinement. In order to reduce the effect of the strong correlation between B factors and occupancy at low resolution, we refined these two parameters separately. Occupancies were first refined during TLS-only refinement, with each lead ion incorporated into the same TLS group as its contacting acid residue. Individual B factors were then refined with occupancies fixed.

Refinement in *PHENIX* is typically accomplished by first refining a TLS-only B -factor model for 5–10 rounds, with the starting coordinates held fixed. Coordinates are then refined to convergence with optimization of X-ray/stereochemistry weight and with torsion-angle restraints ($\sigma = 1\text{--}2^\circ$) using the starting model as a reference. While further refinement with the use of restrained individual B factors may be warranted once the quality of the model is sufficient, this should be used with care (Merritt, 2012).

Paired refinement of C4b was carried out using the *PDB_REDO* web server (Joosten *et al.*, 2014), which implements a modified version of the protocol introduced by Karplus & Diederichs (2012), in which the structure is refined at multiple resolution cutoffs, with the limit defined when adding an extra shell reduces the R factors at the previous resolution limit. We performed this analysis on both the original published coordinates and a partially refined model in which the register errors were corrected and the structure as a whole was equilibrated in the MDFF environment but prior to complete rebuilding. In the former case the original 4.2 Å resolution cutoff was recommended, whereas the corrected and equilibrated model supported the extension of the data set to the limit of data collection at 3.5 Å.

3. Results

3.1. Overview of changes to the C4 structure

An overview of the changes made to the C4 structure (PDB entry 4fxk) is shown in Fig. 2, and is for the most part representative of the changes made to C4–MASP-2 and C4b as well. The largest individual changes are of course the register shifts in the MG3 and CUB domains (marked with an asterisk and a dagger, respectively). A video recapitulating the task of remodelling the CUB domain is available as Supplementary Movie S1. Numerous smaller-scale corrections were made to all three structures, as reflected in the significantly improved clashscore, Ramachandran and rotamer statistics (Table 1). In the case of the C4 structure (PDB entry 4fxk) we were also able to model the O-linked GalNAc at Thr1244 and four trimethyllead ions into $+3\sigma$ $mF_o - DF_c$ density, the latter present at occupancies too low to yield an obvious anomalous signal in the original experiment (Kidmose *et al.*, 2012). The highest occupancy trimethyllead ion forms a stabilizing bridge

Figure 2

Overview of the C4 structure and the key corrections made in this study. Register errors affecting two β -strands and a poorly ordered loop in the MG3 domain (marked with an asterisk) and four β -strands in the CUB domain (marked with a dagger) were corrected with shifts of 1–3 residues. All $\text{C}\alpha$ atoms which shifted by more than 2.5 Å are shown in sphere representation and coloured according to the distance moved. Additionally, four low-occupancy trimethyllead ions were resolved (black spheres and arrowheads) as well as the O-linked *N*-acetylgalactosamine (arrow; obscured behind protein in this image).

Table 1
Structure solution and refinement.

Values in parentheses are for the outer shell. There is twofold noncrystallographic symmetry in the C4–MASP-2 and C4b structures.

	C4	C4–MASP-2	C4b
Resolution range (Å)	45.88–3.60 (3.73–3.60)	48.97–3.75 (3.86–3.75)	49.38–3.50 (3.63–3.50)
$R_{\text{r.i.m.}}^{\dagger\dagger}$ (%)			53.1 (338.8)
Completeness (%)	99.6 (99.9)	98.4 (97.1)	99.0 (99.5)
Multiplicity [†] (%)			7.0 (7.2)
$\langle I/\sigma(I) \rangle^{\dagger\dagger}$			4.55 (0.65)
$CC_{1/2}^{\dagger\dagger}$ (%)			94.9 (15.2)
No. of reflections, working set	26947 (2655)	56606 (5548)	60234 (6041)
No. of reflections, test set	1348 (133)	1727 (197)	2014 (200)
Final R_{cryst}	0.215 (0.289)	0.212 (0.305)	0.248 (0.342) [0.221§]
Final R_{free}	0.263 (0.364)	0.268 (0.342)	0.294 (0.361) [0.264§]
No. of non-H atoms			
Protein	12684	31980	24256
Ligand	207	234	56
Total	12891	32214	24312
R.m.s. deviations			
Bonds (Å)	0.002	0.003	0.003
Angles (°)	0.62	0.51	0.83
Average B factors (Å ²)			
Protein	132.6	130.2	136.5
Ligand	186.8	166.1	104.1
Ramachandran plot			
Most favoured (%)	95.2	95.6	94.6
Allowed (%)	4.6	4.2	5.3
Rotamer outliers (%)	2.7	1.7	2.3
Clashscore	1.6	2.8	1.4
<i>MolProbity</i> score	1.58	1.55	1.52
TLS groups	16	38	24

[†] Merging statistics for C4b are reported owing to extension of the resolution from 4.2 to 3.5 Å. Statistics for other structures are as published previously (Kidmose *et al.*, 2012). [‡] Resolution cutoffs were determined by paired refinement in *PDB_REDO* (Joosten *et al.*, 2014) after partial rebuilding. $CC_{1/2}$ and $\langle I/\sigma(I) \rangle$ in the outer shell are consistent with the cutoffs suggested in Diederichs & Karplus (2013). $\langle I/\sigma(I) \rangle$ falls below 2.0 between 4.21 and 4.06 Å. [§] Calculated using a high-resolution cutoff of 4.2 Å.

between acid residues on crystallographically related molecules, consistent with the original observation that soaking of the crystal in trimethyllead chloride increased the resolution and decreased the anisotropy. While very few high-resolution examples containing this ion exist in the PDB, the geometry at this site is very similar to that seen in the 2.15 Å resolution *Escherichia coli* maltose-*O*-acetyltransferase structure (PDB entry 1ocx; Lo Leggio *et al.*, 2003), where trimethyllead also bridges two acidic residues. The other three trimethyllead ions are bound to one acid residue each, with Pb–O bond distances consistent with previous observations (Holden & Rayment, 1991; Biou *et al.*, 1995). In the case of C4b, paired refinement of the revised model using the *PDB_REDO* server (Joosten *et al.*, 2014) supported extension of the resolution cutoff from 4.2 to 3.5 Å. When assessed against data truncated to the original 4.2 Å resolution, the R_{free} of our final C4b model was reduced by 0.9% compared with that of PDB entry 4xam, while R_{work} was increased by 0.4%. The $CC_{1/2}$ and $\langle I/\sigma(I) \rangle$ in the outermost shell were 0.152 and 0.65, respectively, in good agreement with the cutoffs for these parameters recently suggested by Diederichs & Karplus (2013).

3.2. Rebuilding of the CUB domain

The CUB domain is discontinuous, with two stretches connected by the TE domain, and comprises residues 935–982 and 1321–1391. Initial remodelling *via* iMDFF was carried out

primarily by inspection and interactive adjustment of the regions surrounding individual Ramachandran outliers. Typically, a 7–12-residue stretch centred on the outlying residue combined with the nearest-neighbouring residues were mobilized as iMDFF simulations. In most situations this represents a good compromise between providing sufficient scope to identify and fix most problems while keeping the simulation small enough to allow sufficient speed for real-time interaction. A register error affecting residues 951–994 in the first half of the CUB domain and the beginning of the TE domain (Fig. 3) was identified during the inspection of Asp984–Pro985. While this was not flagged as an outlier in the original coordinates, a persistent Ramachandran outlier quickly developed here upon equilibration in iMDFF. While the local difference map provided very little guidance, further inspection revealed significant excess $2mF_o - DF_c$ density near Glu992 (suggesting an unfolding of what had been assigned as an ideal α -helix) and apparent over-packing of the loop in the vicinity of Leu951. While this proved sufficient to confidently assign a two-residue C-terminal-

directed register shift in this stretch, we later noted that the simple act of re-refining the original coordinates with a simplified TLS-only B -factor model dramatically increased the appearance of $\pm 3\sigma$ $mF_o - DF_c$ density around the erroneously assigned residues (Fig. 3*b*), suggesting that over-parameterization of the original B -factor model played a key role in obscuring the error here. Further inspection of this region led us to also assign register shifts in the adjacent residues 1347–1375 in the second half of the CUB domain. Both register shifts in the CUB domain were subsequently confirmed in the structures of both the C4–MASP-2 complex and in C4b, where the CUB domain is placed in a completely new environment as a result of the conformational change occurring upon C4 cleavage and release of the small C4a fragment. Despite the slightly lower quality crystal data for these two structures, the presence of twofold noncrystallographic symmetry in each significantly improved confidence in the revised register assignments.

3.3. Rebuilding of the MG3 domain

The MG3 domain formed by residues 239–361 was originally built in the 3.6 Å electron density for C4, and is associated with the weakest density of all domains in this structure (Fig. 4). A potential single-residue register error was first noted during a systematic end-to-end investigation of the

Figure 3

Correction of register errors affecting four β -strands in the CUB domain. In this and all further figures, blue wireframe = $2mF_o - DF_c$ map contoured at 0.8σ , blue transparent surface = $2mF_o - DF_c$ ($B_{\text{sharp}} = -80 \text{ \AA}^2$) contoured at 1.5σ , yellow and red translucent surfaces = $mF_o - DF_c$ contoured at $+3\sigma$ and -3σ , respectively. $2mF_o - DF_c$ maps are masked to within $3 \text{ \AA}$ of mobile atoms, while $mF_o - DF_c$ maps are masked at $8 \text{ \AA}$. C^α atoms are coloured according to Ramachandran status (red, outliers; white, generously allowed; blue, preferred). (a, b, c) Detail of the junction between the CUB and TE domains. (a) The register error shown by excess density near Glu992 (*) was initially identified during investigation of the persistent Ramachandran outlier that appeared at Asp984 upon energy minimization (arrow). (b) While this domain was successfully corrected using iMDFF with maps generated directly from the deposited PDB file (with individual B factors), we later noted that re-refinement using a TLS-only B -factor strategy significantly strengthens the evidence of error in the $mF_o - DF_c$ map. (c) Corrected coordinates refined using the same TLS-only B -factor strategy reveal a two-residue out-folding of what was modelled as an ideal α -helix. (d, e) Overview of $mF_o - DF_c$ density (from TLS-only refinement) surrounding the out-of-register stretches before (d) and after (e) correction. †, ‡ and ††, ‡‡ denote the over-packed and under-packed ends, respectively, while the 951–994 and 1347–1375 stretches are shown as orange and green cartoons, respectively. Arrowheads indicate Ramachandran outliers that appeared upon equilibration in iMDFF, while arrows indicate the approximate shift in individual C^α atoms. Remodelling of this domain in iMDFF is also demonstrated in Supplementary Movie S1.

complete structure in iMDFF, but was tentative and based on somewhat circumstantial evidence. Specifically, β -strand residues 350–357 appeared likely to be shifted by one residue, primarily on the basis of a slight under-filling of density N-terminal to Trp360 and a lack of density associated with Glu356 despite its apparently constrained environment. This putative error was confirmed upon correction of the same segment, with the further addition of a one-residue shift in the preceding strand, in the C4-MASP-2 complex, where the density associated with the MG3 domain was substantially stronger. The entire register shift was finally also verified in the C4b structure. Similar to the case of the CUB domain, we later noted that resetting B factors and refining a TLS-only

model (with or without further positional refinement) yielded strong $\pm 3\sigma$ $mF_o - DF_c$ density associated with these errors. This suggests that the use of a more conservative B -factor model would in this case have allowed direct, confident identification and correction of these register errors even given the very weak electron density for this domain in the C4 structure.

4. Discussion

4.1. Implications for the biology of complement C4

The improved structure of the CUB domain is of significant relevance to C4 biology, since this domain is degraded by

Figure 4

Correction of two β -strands in the MG3 domain. In the C4 structure the density here is much weaker than in the CUB domain, and maps generated from the original file (a) gave little evidence of error. A tentative correction was made to the strand containing Glu356 based upon slight under-filling of density N-terminal to Trp360 and an implausible lack of density corresponding to Glu356. No evidence of error was seen in the preceding strand at this stage. (b) We later noted that refining a TLS-only B -factor model against the original (b) or iMDFF energy-minimized (c) coordinates yielded strong evidence in favour of our corrections in the $mF_o - DF_c$ maps. *, **, †, †† and ‡, ‡‡ indicate negative and positive difference density consistent with shifts in Tyr339, Glu356 and Thr358, respectively. The positive difference at ‡‡ is also partially explained by the side chain of Arg337. (d) Final coordinates in the C4 structure. This domain is much more clearly resolved in the C4-MASP-2 (e) and C4b (f) structures (final coordinates and maps), presumably owing to the conformational change occurring in C4 upon MASP2 binding and substantial contacts with the neighbouring MG8 domain in C4b.

proteolytic processing *in vivo* first at Arg1336 and subsequently at Arg956 by protease factor I (FI) assisted by the regulators of complement activation (RCAs) MCP, C4bBP and CR1 (Bajic *et al.*, 2015). The products of these proteolytic degradation reactions are known as iC4b and C4dg, which are both unable to bind zymogen C2 and subsequently support formation of the C3 convertase. As the RCAs are preferentially associated with host cells compared with pathogens, this efficiently limits complement activation on host tissue while allowing it to proceed on pathogens. The RCA proteins share a modular structure based on linearly arranged CCP modules. No detailed information is available regarding how the regulators recognize C4b and promote its degradation by FI, but it is generally believed that this proceeds through a general mechanism very similar to that for degradation by FI of C3b, the functional homologue of C4b. The structure of C3b in complex with four CCP domains of the RCA factor H (FH) in combination with the structure of FI has led to the formulation of a model in which FH provides a platform for orienting FI correctly with respect to the C3b CUB domain for subsequent proteolytic cleavage reactions (Roversi *et al.*, 2011; Wu *et al.*, 2009). If this also applies to C4b degradation, some of the residues within the CUB domain that have been relocated in our improved structures are important. Firstly, the loop containing Arg1349 is likely to be close to the RCA molecule and the subsequent residues up to Glu1355 are facing the C4b MG2 domain and hence are important for the CUB-MG2 contact, which may be crucial for efficient binding of the RCA. The cleavage site itself at Arg1356 in the C4b-RCA-FI complex is close to the two reconstructed loops

Gly962–Asp973 and Ser1361–Lys1365. Hence, an attempt to model the C4b–RCA–FI complex would be more reliable if the updated C4b model were used instead of the old model with register errors in the regions mentioned above.

We have very recently described structures of the pro-convertase C4b–C2 complex and the active convertase C4b–C2a based on rigid-body refinement against SAXS scattering data (Mortensen *et al.*, 2016). In this work the former C4b structure was used as part of the input models, but the conclusions in this study are not affected by the changes we have introduced in C4b presented here, primarily owing to the low resolution of the SAXS data and the fact that the reconstructed CUB and MG3 domains do not interact with either C2a or C2 in the models resulting from rigid-body refinement against the SAXS data. Likewise, our SAXS-based solution structure of C4b (Mortensen *et al.*, 2015) is unlikely to change significantly if the altered C4b model is used as input for rigid-body refinement.

4.2. Implications for model building and refinement methods

The register errors described above each originated in the initial molecular replacement model for C4, which was generated from the C3 structure (PDB entry 2a74; Janssen *et al.*, 2005) using *MODELLER* (Eswar *et al.*, 2008). In each case, local sequence identity between C3 and C4 is poor in the regions that contained register errors (Fig. 5) despite the very similar structure. In addition, there is sufficient ambiguity in side-chain properties and shape (*e.g.* ³⁴⁰VAAAIIE) to preclude confident *ab initio* assignment of the β -strand register in a low-resolution map. Compounding the challenge, the C4 used for crystallization was a mixture of C4A and C4B

purified from expired, pooled human plasma from a local hospital, leading to the presence of variant sequences and the potential for variations in post-translational modifications. When combined with the fact that the maps derived from the original refinements showed surprisingly few ‘red flags’ pointing to serious errors in these locations, it is perhaps not surprising that these register errors went unnoticed. In summary, the example of the C4 structures emphasizes that when using homology modelling to generate the starting model, regions with low sequence conservation, the absence of large side chains and a lack of anomalous signal from heavy atoms to validate the assigned register should be investigated carefully and repeatedly throughout iterative cycles of model building and refinement. In this context, the ability of iMDFF to explicitly model local physical interactions and rapidly evaluate different register scenarios appears to be particularly valuable, as previously demonstrated in the iMDFF-aided rebuilding of the insulin receptor ectodomain (Croll *et al.*, 2016) and further demonstrated in Supplementary Movie S1.

Our observations during the course of this re-refinement suggest various avenues to reduce the incidence of such problems in future. We stress that the majority of these suggestions are by no means new, and are provided primarily as a reinforcement of the extensive existing literature on this subject.

4.2.1. Reduce the conformational degrees of freedom. Like most structures at these resolutions, the three structures discussed here suffer from a very low observation-to-parameter ratio, with only 1.8–2.5 working reflections per non-H atom. This is mitigated somewhat in the case of C4–MASP-2 and C4b by the presence of twofold NCS, but nevertheless overfitting during refinement in such situations is a constant

danger. As has been thoroughly documented elsewhere (Kleywegt & Jones, 1997; Merritt, 2012) this can be reduced by either reducing the number of fitting parameters or introducing extra ‘observations’ in the form of knowledge-based restraints, or some combination of both.

Existing model-building packages such as *Coot* (Emsley & Cowtan, 2004), *O* (Jones *et al.*, 1990) or *MAIN* (Turk, 2013) were designed to be run on the desktop computers available in the 1990s or early 2000s, and their design reflects the limited computational resources of these machines. Calculation of nonbonded interactions in particular is a very computationally demanding task, and is heavily simplified or entirely absent in these packages. User interaction is typically limited to rigid-body translation

Figure 5

The three stretches containing register errors have poor identity to C3, the template used to generate the initial model in *MODELLER* (Eswar *et al.*, 2008). For each alignment, the consensus motif of mammalian C3 species is shown above with the human C4 sequence aligned according to the original and final structures below. For C4 sequences, residues identical to the most common C3 variant are green, residues identical to at least one C3 variant are blue, residues with similar properties to the C3 equivalent are orange and the remainder are grey. Sine waves and arrows under each alignment indicate the extent of α -helices and β -sheets, respectively, before and after revision. The placement of Pro985 as the second residue entering an α -helix is a common motif (Kim & Kang, 1999).

and rotation of fragments (*e.g.* *via* selection of specific rotamers) and/or simple ‘rubberband’ stretching and relaxation into the nearest density while surrounding atoms are ignored. It is a routine occurrence for the crystallographer to be put in a position where in order to fix one error he or she must introduce new errors in the form of atomic clashes or unphysical geometry and rely on automated regularization and refinement procedures to reduce these. While these methods allow the building of high-quality models into atomic resolution data, as the resolution degrades the task becomes increasingly difficult as the density no longer provides sufficient guidance to impose realistic stereochemistry. The scale of the challenge is reflected by the generally poor stereochemical quality seen in published low-resolution structures (Fig. 1).

The concept of interactive molecular dynamics has a long history, having been first introduced by Cyrus Levinthal in an article for *Scientific American* in 1966 (Levinthal, 1966), and was added to *VMD* in 2001 (Stone *et al.*, 2001). However, it is only within the last decade that it has become possible for individual, reasonably priced workstations to run simulations involving thousands of atoms at speeds sufficient for real-time interaction. By explicitly modelling van der Waals and electrostatic interactions, such simulations add a large and very powerful set of extra restraints to the modelling task, excluding a substantial proportion of unphysical degrees of freedom from the space that the crystallographer must search. As atoms are manipulated by the application of external forces the simulation continually acts to reimpose physicality upon the surrounding structure, such that the modelling task becomes akin to working with a real-world model of a protein. Many tasks (such as the adjustment of rotamers or the flipping of individual peptide bonds) which in existing packages involve multiple menu-driven steps to choose, apply and then settle the revised conformation are achieved using iMDFF in seconds by simply pulling briefly on a key atom. In this way, the need for rigid-body rearrangements is removed for all but the most extreme problems (*e.g.* those which would require bonds to pass through each other to fix). Furthermore, thermal equilibration in a realistic molecular-dynamics environment tends to significantly improve backbone and rotamer conformation and packing even in the absence of user intervention, as residues that were already ‘almost correct’ settle into more energetically favourable states. This is identical to the behaviour of the conceptually similar but non-interactive xMDFF (McGreevy *et al.*, 2014), but we stress that user intervention (as in iMDFF) and/or more sophisticated search algorithms (DiMaio *et al.*, 2013; LuCore *et al.*, 2015) remain necessary to efficiently handle conformational adjustments involving substantial energy barriers.

4.2.2. Focus on stereochemical quality and real-space correlation before minimizing residuals. As well as increasing the complexity of the space through which the crystallographer must search while model building, we argue that the introduction of atomic clashes and unnatural geometry allowed by current model-building tools also inadvertently increases the number of degrees of freedom available to the refinement algorithm. While scores for common stereo-

chemical indicators such as clashes, bond lengths and angles *etc.* are included in the target for all modern refinement programs, these remain primarily local searches prone to settling into deep local minima. The essential problem here is that there are far more ways to improve the *R* factors while reducing (for example) a clashscore from 70 to 69 than to do so while not introducing clashes into a clash-free model. This problem is compounded at low resolution, particularly in more disordered regions, where it becomes less and less reasonable to expect that the best fit to the data would correspond to a natural structure. Thus, it is to be expected that models hand-built into low-resolution data *via* traditional methods will be prone to overfitting and that correction of their geometry will often involve substantial initial increases in R_{free} . This was our experience with all three of the structures discussed here, where initial rounds of iMDFF led to R_{free} values 2–6% higher than for the original structures. While further rebuilding and refinement eventually reduced the R_{free} values of C4 and C4b below those of the original models, that of C4-MASP-2 remains almost 3% higher than the original despite significant improvement in stereochemistry and map quality. We note, however, that in all three cases the $R_{\text{free}}-R_{\text{work}}$ gap is significantly decreased, indicating an overall reduction in overfitting. Whilst increasing the X-ray *versus* stereochemistry weighting during refinement was able to further reduce R_{free} by 1–4%, this came at the expense of a large and unjustifiable degradation in stereochemical quality.

4.2.3. Simplify the *B*-factor model. As demonstrated by Merritt (2012), the use of individual *B* factors in low-resolution refinement is not necessarily supportable even when it leads to a lower R_{free} value. This is justified statistically based on answering the question of whether the decrease in R_{free} is sufficient to justify the increased degrees of freedom of the more complex model (the Hamilton *R*-factor ratio test). Unfortunately, as described in Merritt (2012), in practice this ratio is nearly impossible to measure accurately, since the number of degrees of freedom is poorly defined, particularly in the presence of distance-based and/or prior information-based restraints. In contrast to this work, the *B*-factor restraint scheme implemented in *PHENIX* is considered to allow the refinement of pseudo-individual *B* factors at resolutions as low as 4.2 Å (Afonine *et al.*, 2012). However, our experience indicates that the choice of TLS-only or individual *B*-factor models may also be justified by asking a simpler question: how does the introduction of individual *B* factors affect the interpretability of the maps? To answer this question one must keep in mind that a *B* factor can be a meaningful representation of atomic displacement if and only if the atom it applies to is placed at the true centroid of its distribution of positions within the crystal. For atoms that are significantly misplaced, the *B* factor becomes little more than a ‘fudge factor’, refinement of which can only serve to reduce the visibility of the error in the resulting maps. Large numbers of such errors may combine to introduce artefacts elsewhere throughout the structure and potentially obscure or obfuscate the structure of more weakly resolved regions. This is consistent with our experience here, where $mF_o - DF_c$ maps

arising from TLS-only refinement clearly revealed numerous positional errors which were not apparent in maps arising from refinement with individual B factors. We therefore suggest that best practice when dealing with low-resolution data should be to perform the majority of model building and refinement using a physically reasonable TLS-only model (here we have defined one TLS group per protein domain, with the larger N-linked glycans assigned to their own TLS groups). Once the model has converged under this scheme it may be beneficial to add restrained individual B -factor refinement, but the results should be carefully inspected to see whether this adds further useful information. In the case of C4, for example, we found that addition of individual B -factor refinement to a model previously refined to a *MolProbity* score of <1.5 with TLS only led to a decrease in R_{free} , a slight improvement in geometry and the addition of density to the maps corresponding to mannose antenna residues on two N-linked glycans.

4.2.4. Re-refine your reference models. One of the best strategies for the improvement of low-resolution models remains the use of conformational restraints to higher resolution reference models where these are available. Such is the case for the C4-MASP-2 complex, where MASP-2 was restrained to the two existing fragment structures [PDB entries 1q3x (2.23 Å resolution; Harmat *et al.*, 2004) and 1zjk (2.18 Å; Gál *et al.*, 2005)]. It is typical for such reference models to be used as they are, effectively carrying through any residual errors into the lower resolution structure. Re-refinement of older models to modern standards would be expected to provide the best possible outcome, but represents a substantial time commitment using traditional methods and is rarely performed. Automated rebuilding and re-refinement pipelines such as *PDB_REDO* (Joosten *et al.*, 2012) may offer some improvement, but results remain variable and the gold standard arguably remains direct inspection and correction against the maps. Using 1–2 rounds of iMDFF interspersed with refinement in *PHENIX* and spending no more than 2 h of interactive time on each structure, we were able to improve PDB entry 1q3x from its original *MolProbity* score of 2.34 and R_{free} of 0.224 to 1.35 and 0.220, respectively, and PDB entry 1zjk from 2.08 and 0.253 to 1.34 and 0.226, respectively. While most improvements were quite minor, we eliminated ten Ramachandran outliers and identified a previously missed *cis*-peptide bond between Ser312 and Pro313. Refinement statistics are provided in Supplementary Table S1.

4.3. Potential synergy between iMDFF and other modelling/validation tools

The design of iMDFF has been guided by the twin hypotheses that (i) high-quality and information-rich visualization will improve the ability of a user to directly identify structural errors and (ii) that a physically realistic modelling environment will both ease the identification of solutions and prevent many common errors from occurring in the first place. While various sophisticated validation tools exist [e.g. *MolProbity* (Chen *et al.*, 2010) or *WHAT_CHECK* (Hooft *et*

al., 1996)] to flag sites of likely error, in general these do not in themselves suggest solutions to the (often daunting) lists of potential issues that they produce. Some individual residues found in the out-of-register regions in the original C4 structure appeared as outliers in reports generated by both servers, but did not clearly stand out as more serious than more localized errors elsewhere. On the other hand, output from the *ERRAT* server (Colovos & Yeates, 1993; which scores nine-residue sliding windows according to nonbonded interactions with their neighbours, *via* statistical comparison to high-resolution exemplars) scored portions of each out-of-register stretch above its 99% exclusion cutoff.

Automated refinement/rebuilding pipelines as implemented in, for example, *PDB_REDO* (Joosten *et al.*, 2012) or *PHENIX* (Adams *et al.*, 2010) can often resolve many local-scale errors, but are currently incapable of handling larger errors in conformation or register at low resolutions. Fragment-based algorithms such as *Rosetta* refinement (DiMaio *et al.*, 2013), on the other hand, can often find and correct register errors in well defined regions, but it remains the case that no automated routine is perfect at the resolution of the C4 structures. We therefore believe that direct visual inspection and manual adjustment of models will remain necessary for the foreseeable future.

In this study, we have deliberately avoided the use of any detailed reports other than the Ramachandran plot to highlight potential issues in order to test the limits of what can be performed in the iMDFF environment while guided by 3.5–4 Å resolution experimental maps with minimal reference to external data. However, there is substantial scope for future integration of the iMDFF approach with existing (semi-) automated rebuilding and validation tools. As an example of the former, we have already incorporated the use of automated steering forces to reassign the register of α -helices and β -strands, but the choice of target register remains up to the user. Current automated routines typically involve the use of statistical and real-space correlation methods to score different potential registers, but committing of the new register typically involves a rigid-body adjustment necessitating removal and rebuilding of flanking loops, and conformational searches of surrounding side chains to resolve clashes. Within the iMDFF environment, the highest-scoring register could instead be readily defined as a target for the simulation, with the user making direct adjustments where necessary. In terms of incorporation of validation tools, we have found that real-time visualization of Ramachandran outliers (by mapping the percentile score of each residue to the colour of the C^α atom) dramatically improved the speed at which we could identify and correct errors, avoiding the need to constantly stop, regenerate and read the outlier list. A typical workstation now has sufficient spare computational capacity such that similar real-time visualizations could conceivably be achieved for (for example) rotamer outliers, internal voids or unlikely side-chain placements. Additionally, use of the per-atom forces calculated by the MD environment may provide further tools for error detection, for example by highlighting atoms or regions experiencing unusually high

average forces. This is, however, well beyond the scope of the current study.

5. Conclusions

Low-resolution model building in the absence of higher resolution reference models is a very challenging task, with which even highly experienced practitioners struggle using traditional tools designed for use with high-resolution data. By leveraging the computational and graphical capabilities of modern workstations, iMDFF was designed to provide a model-building environment akin to moulding a real-world physical object, allowing rapid and confident remodelling of large and small errors and significantly improving the ultimate quality of the refined model. Here, we have used iMDFF combined with PHENIX to rebuild and refine three related >3.5 Å resolution structures to the standards typically expected from data with a resolution better than 2 Å. In terms of biology, our results provide a better structural framework for understanding the pivotal function of the C4 protein within the complement system.

Acknowledgements

TC and GRA were supported by Aarhus University Research Foundation. GRA was supported by the Danish Science Research Council for Nature and the Universe. Portions of the iMDFF work used a GPU kindly donated by Nvidia Corporation.

References

- Adams, P. D. *et al.* (2010). *Acta Cryst. D* **66**, 213–221.
- Afonine, P. V., Grosse-Kunstleve, R. W., Echols, N., Headd, J. J., Moriarty, N. W., Mustyakimov, M., Terwilliger, T. C., Urzhumtsev, A., Zwart, P. H. & Adams, P. D. (2012). *Acta Cryst. D* **68**, 352–367.
- Bajic, G., Degn, S. E., Thiel, S. & Andersen, G. R. (2015). *EMBO J.* **34**, 2735–2757.
- Biou, V., Shu, F. & Ramakrishnan, V. (1995). *EMBO J.* **14**, 4056–4064.
- Blanchong, C. A., Chung, E. K., Rupert, K. L., Yang, Y., Yang, Z., Zhou, B., Moulds, J. M. & Yu, C. Y. (2001). *Int. Immunopharmacol.* **1**, 365–392.
- Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S. & Richardson, D. C. (2010). *Acta Cryst. D* **66**, 12–21.
- Colovos, C. & Yeates, T. O. (1993). *Protein Sci.* **2**, 1511–1519.
- Croll, T. I., Smith, B. J., Margetts, M. B., Whittaker, J., Weiss, M. A., Ward, C. W. & Lawrence, M. C. (2016). *Structure*, **24**, 469–476.
- Diederichs, K. & Karplus, P. A. (2013). *Acta Cryst. D* **69**, 1215–1222.
- DiMaio, F., Echols, N., Headd, J. J., Terwilliger, T. C., Adams, P. D. & Baker, D. (2013). *Nature Methods*, **10**, 1102–1104.
- Emsley, P. & Cowtan, K. (2004). *Acta Cryst. D* **60**, 2126–2132.
- Eswar, N., Eramian, D., Webb, B., Shen, M.-Y. & Sali, A. (2008). *Methods Mol. Biol.* **426**, 145–159.
- Gál, P., Harmat, V., Kocsis, A., Bián, T., Barna, L., Ambrus, G., Végh, B., Balczar, J., Sim, R. B., Náray-Szabó, G. & Závodszy, P. (2005). *J. Biol. Chem.* **280**, 33435–33444.
- Halim, A., Rüetschi, U., Larson, G. & Nilsson, J. (2013). *J. Proteome Res.* **12**, 573–584.
- Harmat, V., Gál, P., Kardos, J., Szilágyi, K., Ambrus, G., Végh, B., Náray-Szabó, G. & Závodszy, P. (2004). *J. Mol. Biol.* **342**, 1533–1546.
- Headd, J. J., Echols, N., Afonine, P. V., Grosse-Kunstleve, R. W., Chen, V. B., Moriarty, N. W., Richardson, D. C., Richardson, J. S. & Adams, P. D. (2012). *Acta Cryst. D* **68**, 381–390.
- Holden, H. M. & Rayment, I. (1991). *Arch. Biochem. Biophys.* **291**, 187–194.
- Hooft, R., Vriend, G., Sander, C. & Abola, E. E. (1996). *Nature (London)*, **381**, 272.
- Hortin, G., Sims, H. & Strauss, A. W. (1986). *J. Biol. Chem.* **261**, 1786–1793.
- Huang, J. & MacKerell, A. D. (2013). *J. Comput. Chem.* **34**, 2135–2145.
- Humphrey, W., Dalke, A. & Schulten, K. (1996). *J. Mol. Graph.* **14**, 33–38.
- Janssen, B. J. C., Huizinga, E. G., Raaijmakers, H. C. A., Roos, A., Daha, M. R., Nilsson-Ekdahl, K., Nilsson, B. & Gros, P. (2005). *Nature (London)*, **437**, 505–511.
- Jones, T. A., Bergdoll, M. & Kjeldgaard, M. (1990). *Crystallographic and Modeling Methods in Molecular Design*, edited by C. E. Bugg & S. E. Ealick, pp. 189–199. New York: Springer.
- Joosten, R. P., Joosten, K., Murshudov, G. N. & Perrakis, A. (2012). *Acta Cryst. D* **68**, 484–496.
- Joosten, R. P., Long, F., Murshudov, G. N. & Perrakis, A. (2014). *IUCrJ*, **1**, 213–220.
- Karplus, P. A. & Diederichs, K. (2012). *Science*, **336**, 1030–1033.
- Kidmose, R. T., Laursen, N. S., Dobó, J., Kjaer, T. R., Sirotkina, S., Yatime, L., Sottrup-Jensen, L., Thiel, S., Gál, P. & Andersen, G. R. (2012). *Proc. Natl Acad. Sci. USA*, **109**, 15425–15430.
- Kim, M. K. & Kang, Y. K. (1999). *Protein Sci.* **8**, 1492–1499.
- Kjaer, T. R., Thiel, S. & Andersen, G. R. (2013). *Mol. Immunol.* **56**, 222–231.
- Kleywegt, G. & Jones, T. (1997). *Methods Enzymol.* **277**, 208–230.
- Kojouharova, M., Reid, K. & Gadjeva, M. (2010). *Mol. Immunol.* **47**, 2154–2160.
- Law, S. K. (1983). *Ann. NY Acad. Sci.* **421**, 246–258.
- Law, S. K., Lichtenberg, N. A. & Levine, R. P. (1980). *Proc. Natl Acad. Sci. USA*, **77**, 7194–7198.
- Levinthal, C. (1966). *Sci. Am.* **214**, 42–52.
- Lo Leggio, L., Dal Degan, F., Poulsen, P., Andersen, S. M. & Larsen, S. (2003). *Biochemistry*, **42**, 5225–5235.
- Lovell, S. C., Davis, I. W., Arendall, W. B., de Bakker, P. I., Word, J. M., Prisant, M. G., Richardson, J. S. & Richardson, D. C. (2003). *Proteins*, **50**, 437–450.
- Lovell, S. C., Word, J. M., Richardson, J. S. & Richardson, D. C. (2000). *Proteins*, **40**, 389–408.
- LuCore, S. D., Litman, J. M., Powers, K. T., Gao, S., Lynn, A. M., Tollefson, W. T., Fenn, T. D., Washington, M. T. & Schnieders, M. J. (2015). *Biophys. J.* **109**, 816–826.
- McGreevy, R., Singharoy, A., Li, Q., Zhang, J., Xu, D., Perozo, E. & Schulten, K. (2014). *Acta Cryst. D* **70**, 2344–2355.
- McKern, N. M. *et al.* (2006). *Nature (London)*, **443**, 218–221.
- Merritt, E. A. (2012). *Acta Cryst. D* **68**, 468–477.
- Mortensen, S., Jensen, J. K. & Andersen, G. R. (2016). *J. Biol. Chem.* **291**, 16494–16507.
- Mortensen, S., Kidmose, R. T., Petersen, S. V., Szilágyi, A., Prohászka, Z. & Andersen, G. R. (2015). *J. Immunol.* **194**, 5488–5496.
- Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kalé, L. & Schulten, K. (2005). *J. Comput. Chem.* **26**, 1781–1802.
- Roversi, P., Johnson, S., Caesar, J. J., McLean, F., Leath, K. J., Tsiftoglou, S. A., Morgan, B. P., Harris, C. L., Sim, R. B. & Lea, S. M. (2011). *Proc. Natl Acad. Sci. USA*, **108**, 12839–12844.
- Stone, J. E., Gullingsrud, J. & Schulten, K. (2001). *Proceedings of the 2001 Symposium on Interactive 3D Graphics*, pp. 191–194. doi:10.1145/364338.364398.
- Turk, D. (2013). *Acta Cryst. D* **69**, 1342–1357.
- Velankar, S. *et al.* (2016). *Nucleic Acids Res.* **44**, D385–D395.
- Won, Y. (2012). *J. Phys. Chem. A*, **116**, 11763–11767.
- Wu, J., Wu, Y.-Q., Ricklin, D., Janssen, B. J., Lambris, J. D. & Gros, P. (2009). *Nature Immunol.* **10**, 728–733.

STRUCTURAL
BIOLOGY

Volume 72 (2016)

Supporting information for article:

Re-evaluation of low-resolution crystal structures via interactive molecular-dynamics flexible fitting (iMDFF): a case study in complement C4

Tristan Ian Croll and Gregers Rom Andersen

Table S1 Re-refinement statistics for high-resolution reference models used in C4:MASP-2 refinement

	1q3x	1zjk
Resolution range (Å)	35.58-2.23 (2.31-2.23)	27.97-2.181 (2.259-2.181)
Completeness (%)	95	88
No. of reflections, working set	27014(2359)	18326(924)
No. of reflections, test set	1356(114)	938(45)
Final R_{cryst}	0.1743(0.1991)	0.1901(0.3052)
Final R_{free}	0.2197(0.2634)	0.2260(0.3428)
No. of non-H atoms		
Protein	4885	2991
Ion	2	0
Ligand	42	0
Water	391	104
Total	5320	3095
R.m.s. deviations		
Bonds (Å)	0.003	0.003
Angles (°)	0.59	0.57
Average B factors (Å ²)		
Protein	28.19	51.18
Ion	25.04	N/A
Ligand	28.00	N/A
Water	28.21	45.30
Ramachandran plot		
Most favoured (%)	97	95
Allowed (%)	2.9	4.7
Rotamer outliers (%)	0.58	0.94
Clashscore	4.03	2.39
MolProbity Score	1.35	1.34
Number of TLS groups	0	3