BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 15 2010, pages 1849–1856
doi:10.1093/bioinformatics/btq294

Structural bioinformatics

Advance Access publication June 4, 2010

Exploring the potential of template-based modelling

Braddon K. Lance[1,*], Charlotte M. Deane[2], and Graham R. Wood[1]

[1]Department of Statistics, Macquarie University, North Ryde, Australia and [2]Department of Statistics, University of Oxford, Oxford, UK

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Template-based modelling can approximate the unknown structure of a target protein using an homologous template structure. The core of the resulting prediction then comprises the structural regions conserved between template and target. Target prediction could be improved by rigidly repositioning such single template, structurally conserved fragment regions. The purpose of this article is to quantify the extent to which such improvements are possible and to relate this extent to properties of the target, the template and their alignment.

Results: The improvement in accuracy achievable when rigid fragments from a single template are optimally positioned was calculated using structure pairs from the HOMSTRAD database, as well as CASP7 and CASP8 target/best template pairs. Over the union of the structurally conserved regions, improvements of 0.7 Å in root mean squared deviation (RMSD) and 6% in GDT_HA were commonly observed. A generalized linear model revealed that the extent to which a template can be improved can be predicted using four variables. Templates with the greatest scope for improvement tend to have relatively more fragments, shorter fragments, higher percentage of helical secondary structure and lower sequence identity. Optimal positioning of the template fragments offers the potential for improving loop modelling. These results demonstrate that substantial improvement could be made on many templates if the conserved fragments were to be optimally positioned. They also provide a basis for identifying templates for which modification of fragment positions may yield such improvements.

Contact: braddon.lance@mq.edu.au

Supplementary information : Supplementary data are available at *Bioinformatics* online.

Received on March 7, 2010; revised on May 11, 2010; accepted on June 2, 2010

1 INTRODUCTION

Knowledge of the structure of a protein is crucial to an understanding of its function. Experimental determination of protein structure is far from keeping pace with the discovery of the amino acid sequences of proteins (Levitt, 2007). Thus to further our understanding of the molecular function of organisms using the growing volume of genomic information, it is necessary to be able to accurately predict protein structure from the amino acid sequence.

Template-based modelling (TBM), the most accurate method of predicting protein structure, exploits the structural similarities shared between two related proteins (Cozzetto *et al.*, 2009). As the amino acid sequences of related proteins diverge through evolution, each accumulates insertions, deletions and mutations. These modifications are manifested in shifts and rotations of structural elements, local deformations and changes in side chains (Lesk and Chothia, 1986), while the fold of the protein tends to be maintained. It is this structural similarity of related proteins that is exploited by TBM. Beginning with its amino acid sequence, the unknown, or target structure is predicted by approximating it with the structural features putatively conserved between itself and a known, homologous template structure.

In general, TBM proceeds by first identifying the template structures that best approximate the target, and subsequently attempting to remove any differences that have accumulated between the target and templates. Of the two approaches currently adopted (Bujnicki, 2006), one predicts the target by optimizing restraints on inter-atomic distances and torsion angles, restraints derived in part from the templates (Sali and Blundell, 1993, e.g. Modeller). In the alternative approach, TBM involves the direct transfer of backbone and/or side chain coordinates from the templates to the target structure wherever the target–template sequences are aligned (Bates and Sternberg, 1999; Greer, 1990; Krieger *et al.*, 2009)—it is this last approach with which we are concerned here. The component steps of this approach have been modified little since the work of Browne *et al.* (1969). These are: (i) aligning a template and target sequence; (ii) using the structure from sequence regions conserved between the template and target to predict the target structure; (iii) prediction of non-conserved regions; (iv) side chain addition; and (v) refinement/modification of the resulting predictions.

Current modelling methods are largely successful in finding the initial approximation, that is, in selecting the best template and aligning this with the target. In the two most recent biennial protein structure prediction experiments [CASP7 (Moult *et al.*, 2007) and CASP8 (Cozzetto *et al.*, 2009)], for example, optimal or near-optimal template selection and alignment was achieved for most proteins with readily identifiable templates (Kopp *et al.*, 2007; Kryshtafovych *et al.*, 2009). In spite of this success, the majority of final predictions from TBM methods remain inferior to the direct transfer of coordinates from the best template to the target based on an optimal template–target alignment (Kopp *et al.*, 2007).

The accuracy with which a template approximates a target is directly related to their sequence identity (SI). Templates with residues 30% identical to the target generally approximate the core to around 1.5–2 Å root mean squared deviation (RMSD), and this may improve to <1 Å RMSD when sequence identities rise >50% (Cozzetto and Tramontano, 2005). Even for templates with high SI to

*To whom correspondence should be addressed.

the target, differences in packing and contacts may result in the main-chain conformation of the template and target differing by 'up to several angstroms' (Tress *et al.*, 2005). It is therefore important that TBM be able to predict structures that improve upon the template, and for this it is necessary to modify and refine the backbone (Read and Chavali, 2007; Tramontano and Morea, 2003). Unfortunately, reliable methods of template refinement have so far been elusive.

To completely predict the structure of the target often requires the modelling of regions with no corresponding residues in the template—a consequence of the insertions and deletions that have occurred as proteins evolve. The resulting gaps within the predicted structure must be filled using loop prediction methods (Deane and Blundell, 2001; Rohl *et al.*, 2004). The accuracy of loop prediction is reduced by the distortions present in the unmodified template relative to the target, particularly in the residues adjacent to the gaps in the target–template alignment, known as anchor regions (Deane and Blundell, 2001; Fiser *et al.*, 2000; Lessel and Schomburg, 1999). A further benefit of improving upon the template structure may thus be improved accuracy of anchor regions, which could then yield more accurate prediction of loops.

Modifying a template to better approximate the target is equivalent to adjusting for the structural differences that have accumulated between two homologues. Lesk and Chothia (1986) previously examined the sources of structural differences within the backbone of conserved regions, and identified all differences as a combination of rigid shifts and rotations of the conserved region, and local deformation of the backbone. Using proteins from the eight families known at that time, the way structural differences were manifest between homologues was related to both SI and type of secondary structure. Although a small number of studies such as Baldwin *et al.* (1993) have observed shifts and rotations of structural elements within homologues, and some structure alignment algorithms implement 'flexible' structures (Verbitsky *et al.*, 1999), no other study quantifies the sources of structural variation in the conserved regions of homologues since the analysis of 32 homologues by Lesk and Chothia (1986). Given that more than 50 000 proteins have been experimentally determined in the intervening period, it now seems prudent to revisit this topic.

How much of the structural difference observed between two homologues results from the shifts and rotations of the structural elements, and how much from changes in local conformation? This question is critical for determining the nature of the differences that exist between two homologues. At one extreme, if a large proportion of the differences observed derives from shifts and rotations, then such modifications should be considered prior to exploring local changes. At the other extreme, if all of the differences observed are from changes in local conformation, then a TBM algorithm should focus solely on modifying the local conformation.

Here, we analyse the extent to which differences between the homologous proteins are a consequence of shifts and rotations of the conserved regions. We demonstrate the potential value of optimizing the positioning of template regions conserved within the target, showing that the accuracy with which the conserved fragments within a template structure can approximate the target is predicted by the number and length of structurally conserved regions within the alignment, the α-helical proportion of the target, and the SI. We also show that half of the anchor regions have RMSD to the native conformation decreased by at least 0.45 Å when fragments are shifted optimally, thus providing scope for more accurate loop modelling.

The contributions of the current article are 2-fold. First, the article quantifies the ultimate extent of improvement possible through rotation and translation of structurally conserved regions of a template, so defining the distance to the finish line for such model refinement. Second, the article finds a relationship between the extent of improvement possible and the properties of both the template and the target, so providing initial guidelines for establishing the distance to the finish line for a given template. It remains as future work to provide explicit rules for repositioning fragments, relating the translation and rotation required to local and global properties of the template and target.

2 METHODS

2.1 Datasets

Two datasets presented in the main results were used as a source of target–template pairs, a comprehensive, non-redundant dataset based upon structures within the HOMSTRAD database (Mizuguchi *et al.*, 1998) and a benchmark dataset based upon data from each of the seventh and eighth sessions of CASP (Cozzetto *et al.*, 2009; Moult *et al.*, 2007). (Datasets are provided in the Supplementary Material.) Possible bias arising from the different HOMSTRAD family sizes was avoided by selecting only the pair of proteins from each family with the highest resolution. Furthermore, the HOMSTRAD non-redundant dataset comprised protein pairs for which there was at least one gap in the sequence alignment. HOMSTRAD class was recorded to enable assessment of between-class differences. Families within HOMSTRAD classes represented by fewer than 25 protein pairs in this dataset were excluded, with the final dataset comprising all-α, all-β, $\alpha\beta$, $\alpha+\beta$ and multidomain classes. A 'high resolution' version of this dataset was used for model building, comprising a subset of 232 protein pairs with resolution 2 Å or better.

The CASP dataset comprised all protein domains from each of CASP7 and CASP8 that were defined on the CASP web site as containing a single continuous amino acid sequence. The CASP7 data included all such domains from the TBM targets and their best templates as described by Kopp *et al.* (2007). The CASP8 data included all domains in the TBM and high accuracy TBM categories of CASP (www.predictioncenter.org/casp8). The best template for each CASP8 target was found by identifying the parent structures used by each predictor group, generating a structure alignment between each template and the target using TM-align (Zhang and Skolnick, 2005), and selecting the template with the highest GDT_TS score. The resulting dataset contained 131 models from CASP7 and CASP8 combined (Supplementary Material). A third dataset presented in the Supplementary Material was designed to evaluate how applicable these results were to target–template pairs in which homology was more distant. Three hundred randomly selected pairs of proteins that share homology on the superfamily level (but not on the family level) as classified by SCOP (Murzin *et al.*, 1995) were processed in the same manner as described below for the HOMSTRAD dataset.

2.2 Construction of the fragment-optimized template

Within each protein pair, one protein was arbitrarily selected as 'target', and the remaining protein as 'template'. The method for constructing the datasets analysed in this article is illustrated in Figure 1, and consists of two major steps, target–template alignment and calculation of the 'fragment-optimized template'. These two steps are described below.

2.2.1 Target–template alignment To assess how well the template approximates the target (Fig. 1a) using GDT and RMSD, the best target–template alignment must first be known. The sequence alignment between the

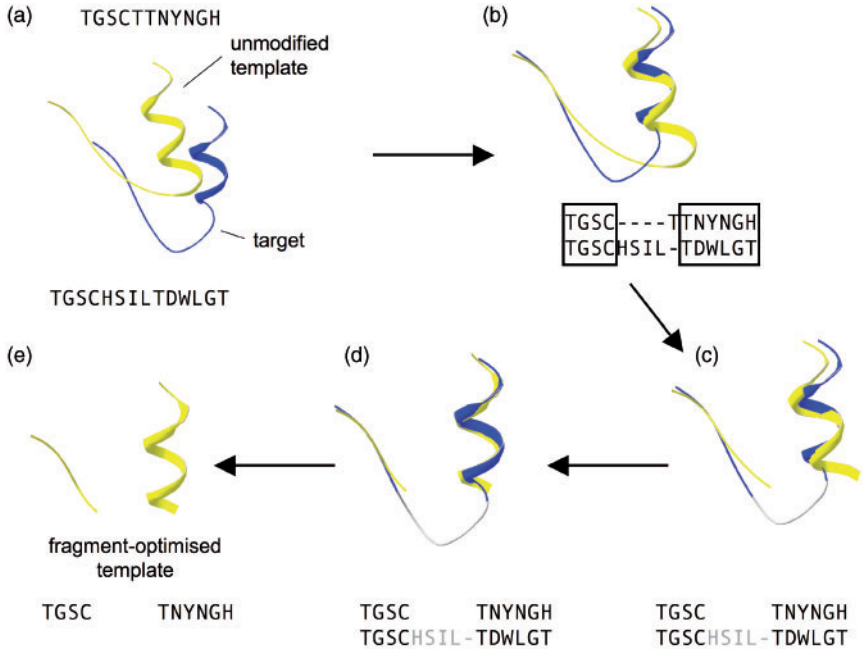

Fig. 1. Optimally modifying the position of template fragments. The method for constructing the dataset: (**a**) the structures within each pair are arbitrarily assigned to target (blue/dark) and template (yellow/light); (**b**) target and template are structurally aligned using TM-align (Zhang and Skolnick, 2005), and regions conserved between target and template in the structural alignment then identified, as indicated by the boxed residues; (**c**) regions not conserved between the target and template are removed from the template, leaving the set of structurally conserved template fragments; (**d**) a fragment-optimized template is created by identifying corresponding regions between template and target within the structural alignment, and individually superposing these conserved template fragments onto the corresponding regions of the target; (**e**) the fragment-optimized template remains as the union of conserved regions, with each conserved region positioned such that they are superposed onto the target.

template and target based upon structural criteria (i.e. structural alignment) was calculated using TM-align (Zhang and Skolnick, 2005). Amino acids within the backbone were defined as 'structurally conserved' between target and template when they were aligned within the structural alignment. Conserved fragments were defined as regions containing four or more contiguous aligned amino acids, indicated in Figure 1b by a 'box' and in Figure 1c as fragments. This definition of structural conservation is a modification of that commonly used in the literature, in which the distance between two structurally conserved residues must also fall below a cut-off (Deane *et al.*, 2001; Hilbert *et al.*, 1993). This modified definition better mimics the real-world task of TBM, where the best sequence alignment may be obtained, but the distance between the C^{α} atoms remains unknown.

Generating the target–template alignment from structure alignment methods such as TM-align corresponds to a 'gold standard', by which alignments made in the absence of knowledge of the target structure may be judged. The aim of this article is to determine the full extent of improvement that is available when adjusting a template to match a target, and so it is consistent with this aim to carry out the alignment with TM-align

To evaluate how the results presented here are affected by a diminished alignment quality more consistent with the task of real-world TBM, the results were recalculated using the sequence-based alignment method CLUSTAL (Larkin *et al.*, 2007). The parameters of the generalized linear model (GLM) presented below were also calculated for this dataset, and are presented in the Supplementary Material.

2.2.2 Calculating the fragment-optimized template When the target and template are globally superposed, the positioning of individual conserved template fragments is not necessarily optimal. To calculate the optimal fragment positions, we performed local least-squares superpositions of the C^{α} atoms from the conserved template fragment onto the corresponding target fragment using Theseus (Theobald and Wuttke, 2006), that is, we optimally positioned each individual conserved fragment to minimize the fragment RMSD (Fig. 1d). The structure with all conserved fragments individually superposed onto the target is shown in Figure 1e, and we refer to this as the fragment-optimized template.

Target–template similarity was assessed using RMSD of all backbone atoms within the union of structurally conserved regions, as well as the 'total score' and 'high accuracy' versions of the global distance test [GDT_TS and GDT_HA (Zemla *et al.*, 1999)]. The amount by which the fragment-optimized template improves upon the unmodified template was quantified by the change in each of RMSD, GDT_HA and GDT_TS, summarized schematically in Figure 2. The change in RMSD, $\Delta\mathrm{RMSD}_a$, was calculated as the RMSD of the unmodified template to target (RMSD_{un}) minus the RMSD of the fragment-optimized template to target (RMSD_{opt}), with both of these quantities calculated for all residues in the conserved fragments (i.e. $\Delta\mathrm{RMSD}_a = \mathrm{RMSD}_{un} - \mathrm{RMSD}_{opt}$). The GDT of the unmodified template was subtracted from the GDT of the fragment-optimized template, so that $\Delta\mathrm{GDT}_a = \mathrm{GDT}_{opt} - \mathrm{GDT}_{un}$.

2.3 Optimal fragment positioning and loop modelling

Gaps in the structure alignment where the target has no corresponding template residues are normally predicted using techniques of loop modelling. Fiser *et al.* (2000) showed that the accuracy with which the conformation of a loop may be predicted is strongly related to the RMSD between the model and target measured over the three residues either side of that loop (the 'anchor' residues). To measure the expected effect of finding the optimal position of individual template fragments on the accuracy of loop modelling, the backbone atom RMSD for the three anchor residues either side of each loop was calculated between both the target and unmodified template, and

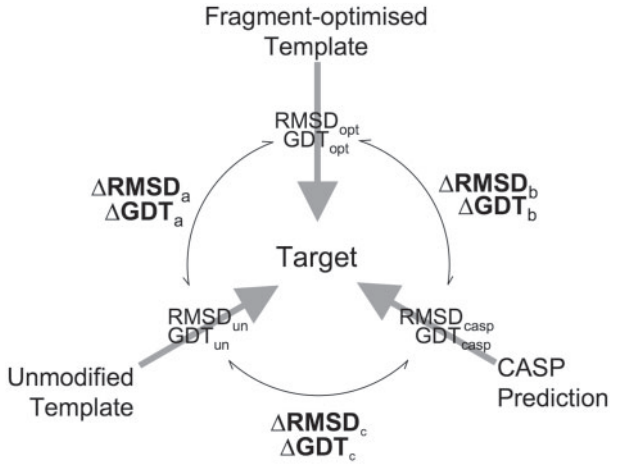

Fig. 2. Three types of template were compared with the target structure: unmodified templates, fragment-optimized templates and the CASP prediction, using two structure similarity metrics, RMSD and GDT. $\Delta RMSD_a$ was calculated as $RMSD_{un} - RMSD_{opt}$ to compare how much better the fragment-optimized template approximates the target than does the unmodified template. For CASP datasets, $\Delta RMSD_b = RMSD_{casp} - RMSD_{opt}$ compares how much better the fragment-optimized template approximates the target than does the best CASP prediction. Similarly, $\Delta RMSD_c = RMSD_{un} - RMSD_{casp}$ compares how much better the best CASP prediction approximated the target than does the unmodified template.

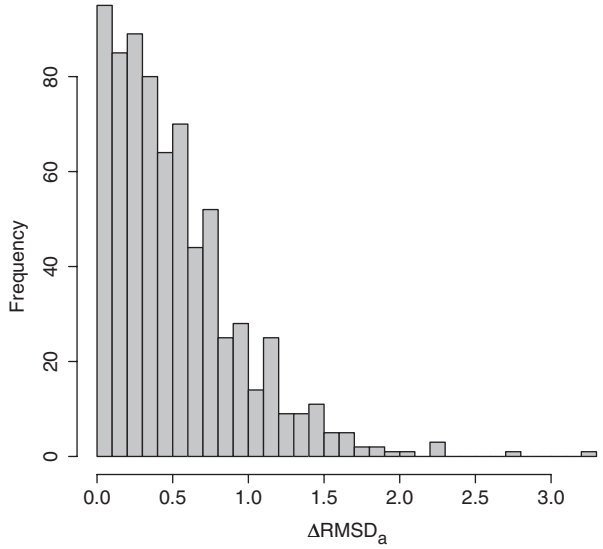

Fig. 3. Distribution of $\Delta RMSD_a$, the difference in RMSD values between the true structure and the template and the true structure and the fragment-optimized template.

the target and fragment-optimized template. The expected change in loop accuracy for each set of anchor residues, ΔRMSD (or anchor ΔRMSD) was calculated as $\Delta RMSD = RMSD_{un} - RMSD_{opt}$, where $RMSD_{un}$ represents the anchor RMSD of the unmodified template to target, and $RMSD_{opt}$ represents anchor RMSD of the fragment-optimized template to target.

2.4 Improvements in structural similarity as a function of other variables

The distribution of improvement in fit of the template to target resulting from fragment optimization ($\Delta RMSD_a$, see Fig. 2) shown in Figure 3 is constrained by zero on the left, and is strongly right skewed. Conditional on the predictor variables used here, $\Delta RMSD_a$ values resemble the gamma distribution (Supplementary Fig. S1). GLMs (McCullagh and Nelder, 1989)) are useful tools for linear modelling where the distribution of the response variable, conditional upon values of the explanatory variables, assumes a distribution that differs from the normal distribution. A GLM with a log link was used to model the $\Delta RMSD_a$ of structure pairs within the HOMSTRAD high-resolution dataset. The predictor variables used were the number of fragments in the alignment, mean fragment length (MFL), the proportion of α-helix in the template as defined by define secondary structure of proteins (DSSP) (Kabsch and Sander, 1983), and SI. Additional variables that were also considered but that did not contribute significantly to the final model included structure class (as defined by HOMSTRAD), target length, template proportion of β-strand, target–template differences in secondary structure composition, proportion of target covered by the template and functions of the difference in side chain volume. Although these variables may in isolation be related to ΔRMSD, they did not explain a significant amount of the remaining variation in ΔRMSD when considered jointly with the four selected variables.

The full model with all variables listed above was simplified by iteratively removing the least significant term with P-value >0.05, resulting in a final model with all terms explaining a significant amount of the variation in ΔRMSD. SI is expressed as a percentage in the text, but is expressed as a value between 0 and 1 for the purposes of modelling. The accuracy of the final model was evaluated by predicting the ΔRMSD for all 3974 template–target pairs in CASP8 (Supplementary Material), and the HOMSTRAD pairs for which the resolution of one or more structures was greater than 2 Å.

3 RESULTS AND DISCUSSION

The structural differences between each homologous target–template pair were analysed by comparing how well each of two templates, the unmodified and 'fragment-optimized' template, approximated the target. The fragment-optimized template is the unmodified template with structurally conserved regions superposed onto the corresponding target region (see Section 2 and Fig. 1 for a full description). Structure comparisons were made using RMSD and GDT scores, as detailed in Section 2.

RMSD captures how similar two structures are in their conserved regions, hence is not 'order dependent', that is, RMSD remains unaffected by which structure within a protein pair is selected as the 'target'. In contrast, GDT captures how well the template approximates the whole of the target—GDT is an order-dependent measure designed for protein structure prediction. GDT is thus affected by both changes in the conserved regions, and insertions in the target relative to the template.

3.1 Differences between homologues and improvements in the template

The shifts and rotations of the structural regions conserved between two homologous proteins account for a substantial amount of their structural difference. Figure 3 shows the distribution of the difference in RMSD, $\Delta RMSD_a$ (Fig. 2), for the HOMSTRAD non-redundant dataset. For 42.7% of pairs, the RMSD between template and target is at least 0.5 Å greater than the RMSD between the fragment-optimized template and the target. In over 12% of pairs this difference is greater than 1 Å.

Using the CASP-favoured measure of GDT confirms that the fragment-optimized template and target are substantially more

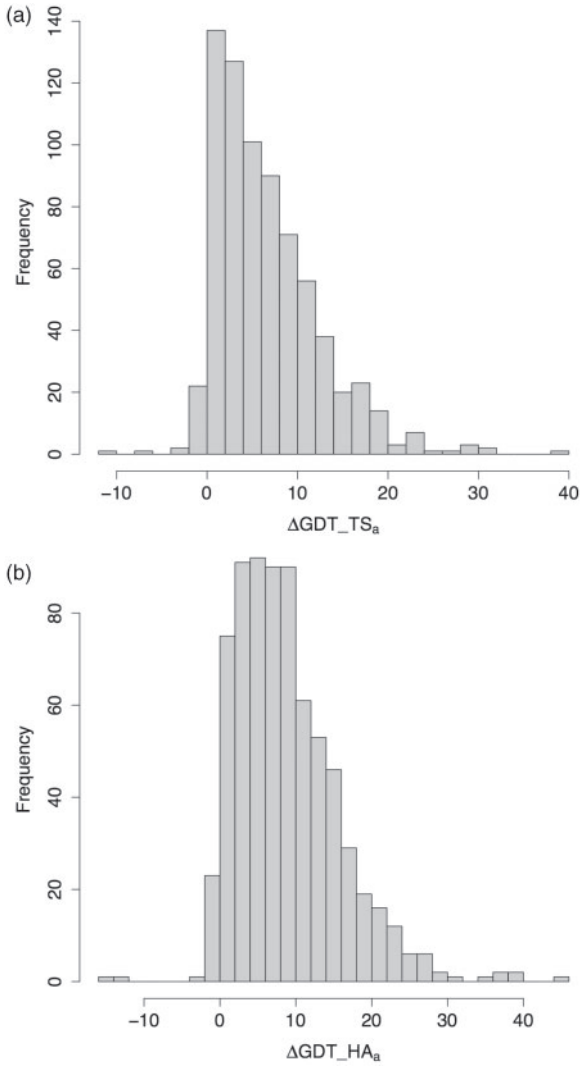

Fig. 4. Histograms of the improvement in GDT after optimal fragment positioning: (**a**) ΔGDT_TS_a; (**b**) ΔGDT_HA_a.

similar than the unmodified template and target. Half of the improvements in the fragment-optimized templates were greater than 5% GDT_TS (Fig. 4a) and 7.7% GDT_HA (Fig. 4b), and one-quarter were greater than 9.7% GDT_TS and 12.6% GDT_HA. Although the bulk of structures were improved as measured by both GDT scores, least squares positioning of template fragments to minimize template RMSD has resulted in a few examples where the fragment-optimized template has a GDT score lower than that of the unmodified template. This apparently anomalous result arises because the kind of differences between structures that are captured by GDT are different from those captured by RMSD, and it is only RMSD that is minimized by least-squares superposition. In cases where the original template fragment was already well positioned relative to the target fragment, very small shifts in the position of some atoms may result in their passing into a more distal distance category (e.g. a single atom moving from 1.99 Å to 2.01 Å from the corresponding target atom would reduce the GDT score). These reductions are best considered 'noise'.

Protein pairs with large values of ΔRMSD and ΔGDT are visible in Figures 3 and 4. These large changes in RMSD upon repositioning of the template fragments arise due to a large number of helices that pack differently in the target and the template. An example of the spatial disparities of multiple helices conserved between the target and template is illustrated in Supplementary Figure S2. The model presented below incorporates the effect of helix packing by including the helical fraction of the template (HFT), and the number of conserved fragments.

3.2 Comparison with CASP predictions

The CASP dataset provides an opportunity to measure the contribution of relative shifts and rotations between the template and target to inaccuracies in current TBM predictions. TBM predictions from CASP7 and CASP8 with best GDT_HA scores and best GDT_TS scores were selected for comparison with both the best template (as listed here and in Kopp *et al.*, 2007) and the fragment-optimized best template. GDT scores for each model relative to the target were calculated based upon the structural alignment from TM-align, and pairwise differences between each of best template, fragment-optimized best template and best CASP model were assessed for significance using paired *t*-tests. The best CASP models have significantly higher GDT_HA (Fig. 5b) and GDT_TS scores than those of the best template ($\overline{\Delta GDT\_HA_c} = 5.9$, $t_{130} = 10.2$, $p < 1e-13$; $\overline{\Delta GDT\_TS_c} = 6.4$, $t_{130} = 10.9$, $P < 1e-12$). The histogram of pairwise differences ΔGDT_HA_b between the fragment-optimized template and best CASP model (Fig. 5b) shows the data generally to the right of the origin, suggesting the fragment-optimized template has a higher GDT_HA and GDT_TS score, although this is not significantly different from zero ($\overline{\Delta GDT\_HA_b} = 1.39$, $t_{130} = 1.9$, $P = 0.055$; $\overline{\Delta GDT\_TS_b} = 0.33$, $t_{130} = 0.52$, $P = 0.60$). Note that ΔGDT_b is calculated over the union of the structurally conserved regions, whereas ΔGDT_c is calculated over the entire protein.

The TBM section of CASP is an assessment of current abilities to approximate a target with one or more homologues, and then account for the structural changes introduced during the course of evolution. Differences that derive from modifications to the positioning of the conserved regions are not yet commonly addressed; accounting for these differences alone would produce gains in model accuracy that are better than those achieved by the best template-based models in CASP.

3.3 Optimal fragment positioning and loop modelling

Loop prediction improves with the accuracy of the anchor residues in the model (Lessel and Schomburg, 1999); we quantify the change in anchor residue accuracy here using the HOMSTRAD non-redundant dataset. The median change in anchor RMSD from unmodified template to target to fragment-optimized template to target is an improvement of 0.45 Å, with 28% of all anchors improved by >1 Å RMSD, as shown in Figure 6. The ΔRMSD of anchor regions shown in Figure 6 is generally greater than the ΔRMSD of the entire structure, as it is anchor regions which change position most following fragment superposition. An extreme example of the large effect that changes in fragment position can have on anchor RMSD is illustrated in Supplementary Figure S3, where repositioning of an SCR in the HOMSTRAD AMP-binding domain improves anchor RMSD by 9.5 Å. Relative to the unmodified template, some anchors

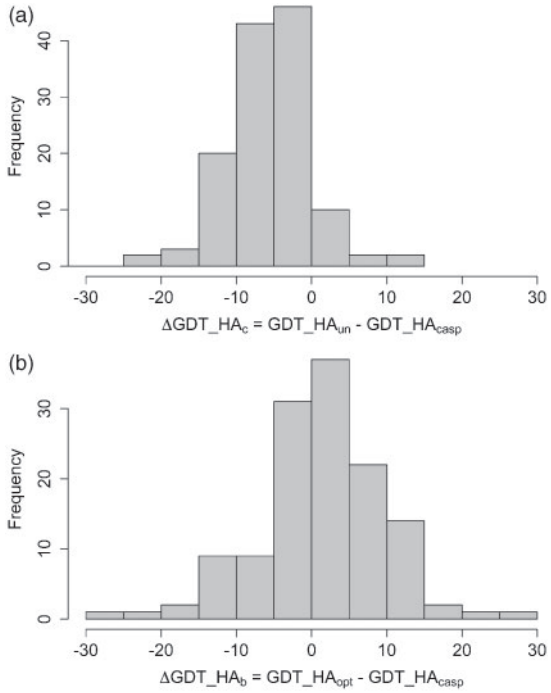

Fig. 5. Difference in GDT_HA of: (**a**) unmodified template to target and best CASP model to target; (**b**) the fragment-optimised template to target and best CASP model to target.

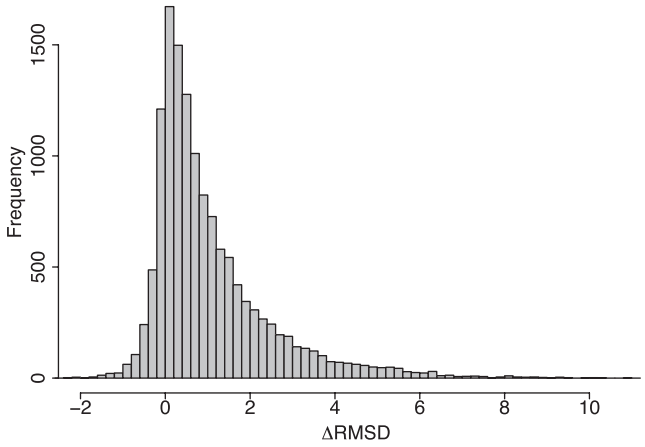

Fig. 6. Distribution of difference in anchor RMSD values between the unmodified template to target, and fragment-optimized template to target.

are positioned less accurately in the fragment-optimized template (these have ΔRMSD below zero in Fig. 6). These reductions in accuracy are almost never greater than 2 Å, and mostly less than 0.5 Å.

Fiser *et al.* (2000) showed that RMSD between predicted and target loops is positively correlated with RMSD between target and template anchor regions. Specifically, for short loops, reductions in anchor RMSD of at least 1 Å are likely to result in reductions in loop RMSD of at least 0.5 Å. Thus, improvements of 0.5 Å or greater can

Table 1. Parameter estimates (fitted model coefficients) for the model of whole-structure $\Delta RMSD_a$

	Estimate	Std error	*t*-value	Pr(> |*t*|)
Intercept	0.9821	0.3771	2.60	0.00981**
log(SI)	−0.2583	0.0912	−2.83	0.00506**
log(MFL)	−0.5851	0.0960	−6.09	4.65e-09***
1/(NF)	−3.6040	0.3183	−11.32	< 2e-16***
HFT	1.1310	0.1983	5.70	3.63e-08***

Significance codes: 0 '***' 0.001 '**'.
Dispersion parameter for gamma family estimated as 0.2654.
Null deviance: 206.938 on 231 df.
Residual deviance: 59.884 on 227 df.
AIC: −147.62.

be expected around a quarter of the time with optimal positioning of conserved fragments.

Better positioning of anchor regions provides a basis for improved loop modelling, so the quality of the prediction on the entire length should be further improved over that currently achieved. Combined with optimal positioning of conserved template fragments, these improvements demonstrate that the ability to optimally shift conserved template fragments could yield improvements in structure prediction better on average than those observed in CASP8.

3.4 Improvements in the template as a function of other variables

Understanding how the characteristics of each protein pair influence the magnitude of the shifts and rotations between two homologues is potentially helpful for TBM; templates that may benefit from such modifications can then be identified.

The GLM for $\Delta RMSD_a$, constructed using the high-resolution HOMSTRAD dataset, is summarized in Table 1. The improvement in the predicted model that results from using the fragment-optimized over the unmodified template varies as a function of the structural similarity of target and template (approximated by SI), alignment properties [the MFL and the number of fragments (NF) in the alignment] and structure type (defined by the HFT, which is highly correlated with the helical fraction of the target). The distributions of these variables are summarised in Supplementary Figure S4, and their relationship with ΔRMSD in Supplementary Figure S5. Specifically, $\Delta RMSD_a$ is modelled multiplicatively as

$$\Delta RMSD_a = 2.670\exp(-3.604NF^{-1})MFL^{-0.585}\exp(1.131HFT)SI^{-0.258}$$

The $\Delta RMSD_a$ for any template–target alignment can be estimated in the absence of the target structure by entering values corresponding to the four variables.

The generality of the model was demonstrated by predicting $\Delta RMSD_a$ for the HOMSTRAD non-redundant dataset, as well as the full CASP8 dataset. The observed versus predicted values are plotted in Figure 7. The HOMSTRAD and CASP8 datasets are seen to be predicted with an accuracy almost identical to that for the data with which the model was built. This consistency indicates that the relationship described between $\Delta RMSD_a$ and the predictors developed here applies generally.

The sensitivity of the parameter estimates to the alignment quality was evaluated by refitting the model using the CLUSTAL

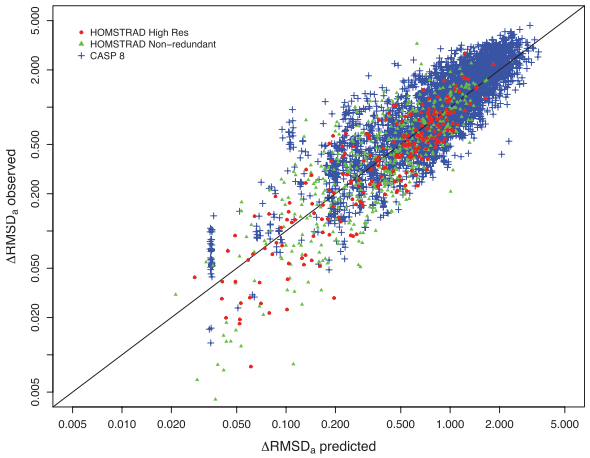

Fig. 7. Observed versus predicted ΔRMSD_a values, showing in red the high-resolution HOMSTRAD data used to build the model, the remaining unbiased HOMSTRAD dataset in green, and the data for all CASP8 targets and templates in blue. Axes use the log scale.

aligned HOMSTRAD pairs. To enable generalization to proteins more distantly homologous, the model was refitted using the SCOP dataset. The recalculated models are presented in Supplementary Tables S1 and S2, and the contrasting parameter estimates discussed qualitatively below.

The two variables of MFL and number of fragments are likely to differ between alignment methods. The predictive model was also re-estimated in the absence of these variables using both the non-redundant HOMSTRAD dataset and the CLUSTAL-aligned HOMSTRAD pairs. These are presented in Supplementary Tables S3 and S4, respectively.

The predictions of ΔRMSD are not uniformly precise for all ΔRMSD values; as ΔRMSD_a increases, the precision with which it can be estimated decreases. The usefulness of this result is evident in that templates for which repositioning of conserved fragments does not yield benefit can be readily identified.

The four predictors are discussed below, in order of importance.

3.4.1 Number of fragments ΔRMSD_a increases with the number of fragments. For example, the ratio of improvements in RMSD as the number of fragments doubles from 5 to 10 is predicted to be $\exp(-3.604(\frac{1}{10}-\frac{1}{5}))\approx 1.4$, or an increase of 40%. The greater the number of fragments in an alignment, the closer the template may be made to approximate the target, the final model being limited by local structure differences in the individual fragments.

For both alignments of poorer quality (represented by the CLUSTAL dataset) and proteins of more distant homology (represented by the SCOP dataset), the effect of the number of fragments on the ΔRMSD is very similar to the effect in the original model (Supplementary Tables S1 and S2).

3.4.2 Mean fragment length ΔRMSD_a increases as the mean conserved fragment length decreases. For example, when the MFL halves, the ratio of improvements in RMSD is $\frac{1}{2}^{-0.585}=1.50$ or an increase of 50%. The shorter the rigid fragments used to approximate the target, the more closely the backbone atoms can be superposed and local structure differences removed.

Note that MFL and NF are functionally related—given the same coverage of target by the template, a shorter fragment length results in a greater number of fragments. The correlation between these variables, however, is weak.

Greater ΔRMSD is achievable in poorer quality alignments at shorter MFLs compared with high-quality alignments (Supplementary Table S1). For high-quality alignments, however, the effect of MFL is essentially unchanged across all levels of homology (Supplementary Table S2).

3.4.3 Fraction of the template that is helical ΔRMSD_a increases with the proportion of residues in α-helices. For example, an increase of 0.1 in the helical fraction yields a predicted increase of $\exp(1.131\times 0.1)\approx 1.12$ or 12% in ΔRMSD_a. Secondary structure within a protein provides regularity, greater for α-helices than for β-sheets. The regularity of helices allows local superposition to be closer, so increasing ΔRMSD_a. The effect of the HFT on ΔRMSD changes little with alignment quality (Supplementary Table S1) and homology (Supplementary Table S2).

The HFT is strongly related to structure class, and so differences in ΔRMSD related to structure class are captured by the HFT (Supplementary Fig. S6).

3.4.4 Sequence identity ΔRMSD_a increases as SI decreases. For example, comparing two templates with 35% and 70% SI to the target, the predicted ΔRMSD is $\frac{1}{2}^{-0.258}\approx 1.20$ or 20% greater for the template with lesser SI. Low SI indicates a poor initial approximation of the target by the template, leaving greater scope for improvement in RMSD through optimal placement of the conserved fragments.

For alignments of poorer quality (here represented by the CLUSTAL dataset), greater improvements in ΔRMSD are predicted (Supplementary Table S1) than when alignments are of higher quality. Analysis of the SCOP dataset (Supplementary Table S2) shows that the effect of SI is consistent even when protein pairs are distantly homologous.

In the absence of variables MFL and NF, SI is still a useful predictor of ΔRMSD for high-quality alignments (Supplementary Table S3). When alignment quality decreases, greater improvements are achievable at lower levels of SI (Supplementary Table S4), just as for the full model presented above.

4 CONCLUSIONS

Pairs of homologous protein structures possess structurally conserved regions that have accumulated differences in structure throughout the process of evolution. Repositioning of such regions is crucial in template-based prediction of protein structure, where one protein is considered a target and the other a template. Such repositioning can be broken into two stages, the first stage being a shift and rotation of the rigid structurally conserved region minimizing target/template RMSD over the region, while the second stage is a deformation bringing the two fragments into coincidence. This article has quantified movements of the first type and related them to properties of the proteins and their alignment.

These results show that a closer approximation of the target (a reduction of the order of 1 Å in RMSD) can often be found via repositioning of conserved structure fragments. Removal of differences between the local backbone conformation of the template

and target necessitates all-atom refinement, and the difficulty of this task is lessened if the fragments can initially be better positioned. The magnitude of the improvement to be gained, and the relatively lower dimensionality of optimally positioning fragments, suggests that the latter is worthy of consideration in TBM.

Conflict of Interest: none declared.

REFERENCES

Baldwin,E. *et al.* (1993) The role of backbone flexibility in the accommodation of variants that repack the core of t4-lysozyme. *Science*, **262**, 1715–1718.

Bates,P. and Sternberg,M.J.E. (1999) Model building by comparison at casp3: using expert knowledge and computer automation. *Proteins*, **S3**, 47–54.

Browne,W. *et al.* (1969) A possible three-dimensional structure of bovine α-lactalbumin based on that of hen's egg-white lysozyme. *J. Mol. Biol.*, **1**, 65–86.

Bujnicki,J. (2006) Protein-structure prediction by recombination of fragments. *ChemBioChem*, **7**, 19–27.

Cozzetto,D. and Tramontano,A. (2005) Relationship between multiple sequence alignments and quality of protein comparative models. *Proteins*, **58**, 151–157.

Cozzetto,D. *et al.* (2009) Evaluation of template-based models in casp8 with standard measures. *Proteins*, **77**, 18–25.

Deane,C. and Blundell,T. (2001) Coda: a combined algorithm for predicting the structurally variable regions of protein models. *Protein Sci.*, **10**, 599–612.

Deane,C. *et al.* (2001) Score: predicting the core of protein models. *Bioinformatics*, **17**, 541–550.

Fiser,A. *et al.* (2000) Modeling of loops in protein structures. *Protein Sci.*, **9**, 1753–1773.

Greer,J. (1990) Comparative modeling methods: Application to the family of the mammalian serine proteases. *Proteins*, **7**, 317–334.

Hilbert,M. *et al.* (1993) Structural relationships of homologous proteins as a fundamental principle in homology modeling. *Proteins*, **17**, 138–151.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Kopp,J. *et al.* (2007) Assessment of casp7 predictions for template-based modeling targets. *Proteins*, **69**, 38–56.

Krieger,E. *et al.* (2009) Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in casp8. *Proteins*, **77**, 114–122.

Kryshtafovych,A. *et al.* (2009) Casp8 results in context of previous experiments. *Proteins*, **77**, 114–122.

Larkin,M.A. *et al.* (2007) Clustal w and clustal x version 2.0. *Bioinformatics*, **23**, 2947–2948.

Lesk,A. and Chothia,C. (1986) The response of protein structures to amino-acid sequence changes. *Philos. Tr. R. Soc. S-A*, **317**, 345–356.

Lessel,U. and Schomburg,D. (1999) Importance of anchor group positioning in protein loop prediction. *Proteins*, **37**, 56–64.

Levitt,M. (2007) Growth of novel protein structural data. *Proc. Natl Acad. Sci. USA*, **104**, 3183–3188.

McCullagh,P. and Nelder,J. (1989) *Generalized Linear Models*. CRC Press, Boca Raton.

Mizuguchi,K. *et al.* (1998) Homstrad: a database of protein structure alignments for homologous families. *Prot. Sci.*, **7**, 2469–2471.

Moult,J. *et al.* (2007) Critical assessment of methods of protein structure prediction - round vii. *Proteins*, **69**, 3–9.

Murzin,A. *et al.* (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Read,R. and Chavali,G. (2007) Assessment of casp7 predictions in the high accuracy template-based modeling category. *Proteins*, **69**, 27–37.

Rohl,C. *et al.* (2004) Modeling structurally variable regions in homologous proteins with rosetta. *Proteins*, **55**, 656–677.

Sali,A. and Blundell,T. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.

Theobald,D. and Wuttke,D. (2006) Theseus: maximum likelihood superpositioning and analysis of macromolecular structures. *Bioinformatics*, **22**, 2171–2172.

Tramontano,A. and Morea,V. (2003) Assessment of homology-based predictions in casp5. *Proteins*, **53**, 352–368.

Tress,M. *et al.* (2005) Assessment of predictions submitted for the casp6 comparative modelling category. *Proteins*, **S7**, 27–45.

Verbitsky,G. *et al.* (1999) Flexible structural comparison allowing hinge-bending, swiveling motions. *Proteins*, **34**, 232–254.

Zemla,A. *et al.* (1999) Processing and analysis of casp3 protein structure predictions. *Proteins*, **3**, 22–29.

Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on tm-score. *Nucleic Acids Res.*, **33**, 2302–2309.