JOURNAL OF THE ROYAL SOCIETY Interface

Assessing the effect of dynamics on the closed-loop protein folding hypothesis

Journal:	*Journal of the Royal Society Interface*
Manuscript ID:	Draft
Article Type:	Research
Date Submitted by the Author:	n/a
Complete List of Authors:	Chintapalli, Sree; University of Essex, School of Biological Sciences Illingworth, Christopher; University of Essex, School of Biological Sciences Upton, Graham; University of Essex, Mathematical Sciences Sacquin-Mora, Sophie; Institut de Biologie Physico-Chimique, Laboratoire de Biochimie Theorique Reeves, Philip; University of Essex, Biological Sciences Mohammedali, Hani; University of Essex, School of Biological Sciences Reynolds, Christopher; University of Essex, Biological Sciences;
Subject:	Computational biology < CROSS-DISCIPLINARY SCIENCES, Biophysics < CROSS-DISCIPLINARY SCIENCES
Keywords:	Closed loops, protein folding, total contact distance, connectivity, hydrophobicity, extended nucleus

Assessing the effect of dynamics on the closed-loop protein folding hypothesis

Sree V. Chintapalli[1a‡], Christopher J.R. Illingworth[1b‡], Graham J. G. Upton[2], Sophie Sacquin-Mora[3], Philip J. Reeves[1], Hani S. Mohammed[1], and Christopher A. Reynolds[1,]*

[1]School of Biological Sciences, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, UK.

[2]Department of Mathematical Sciences, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, UK

[3]Laboratoire de Biochimie Theorique, UPR 9080 CNRS, Institut de Biologie Physico-Chimique, Paris, France.

Authors for correspondence: Christopher A. Reynolds, e-mail: reync@essex.ac.uk

‡SVC and CJRI contributed equally

[a]Current address : Department of Physiology and Membrane Biology, UC Davis School of Medicine, CA 95616, USA.

[b]Current address : Department of Genetics, University of Cambridge CB2 3EH, UK

KEYWORDS Closed Loops, closed loop hypothesis, total contact distance, connectivity, hydrophobicity, protein folding.

RUNNING HEAD: Closed loops and protein folding

The closed loop (loop-n-lock) hypothesis of protein folding suggests that loops of about 25 residues, closed through interactions between the loop ends (locks), play an important role in protein structure. Coarse-grain elastic network simulations, and examination of loop lengths in a diverse set of proteins, each support a bias towards loops of close to 25 residues in length between residues of high stability. Previous studies have established a correlation between total contact distance (TCD), a metric of sequence distances between contacting residues (c.f. contact order), and the log folding rate of a protein. In a set of 43 proteins, we identify an improved correlation (r^2=0.76), when the metric is restricted to residues contacting the locks, compared to the equivalent result when all residues are considered (r^2=0.65). This provides qualified support for the hypothesis, albeit with an increased emphasis upon the importance of a much larger set of residues surrounding the locks. Evidence for a similar sized protein core / extended nucleus (with significant overlap) was obtained from TCD calculations in which residues were successively eliminated according to their hydrophobicity and connectivity, and from molecular dynamics simulations. Our results suggest that while folding is determined by a subset of residues that can be predicted by application of the closed loop hypothesis, the original hypothesis is too simplistic; efficient protein folding is dependent on a considerably larger subset of residues than those involved in lock formation.

1. INTRODUCTION

Amongst the theories on protein folding, Berezovsky et al.'s controversial hypothesis, that the basic protein folding unit is a closed loop (loop-n-lock) with a length of about 25-35 amino acid residues, formed by non-local hydrophobic interactions between the loop ends is of particular interest [1,2]. This hypothesis, which builds on the non-radiative excitation energy transfer measurements of Ittah and Haas [3], is immediately attractive, as it offers the prospect of a molecular level understanding of protein structure and folding, shedding light, for example, on the possible nature of the funnels on folding pathways. The hypothesis can be accommodated into the currently accepted mechanisms of protein folding such as framework, hydrophobic collapse and nucleation-condensation [4-7]. Furthermore, it has potential relevance well beyond the scope of protein folding, for example in matters of protein or drug design.

Current evidence for closed loops (defined in part by a close approach in space of residues some distance apart along the polypeptide chain) comes from several observations, all of which point to a common unit of approximately 25 residues. These observations include a peak in the distribution of the length of protein chain-returns [1,2], a peak in the number of amino acid neighbors as a function of sequence distance [2], the autocorrelation function of hydrophobic residues [8], and of specific hydrophobic tripeptides [9,10], and the presence of minimally disruptive protein fragments, or 'schemas', that can be exchanged without loss of function [11]. Once the locks have been determined (the lock is formed from residues at both ends of the loop), it is observed that hydrophobicity plots show a maximum at the lock residues [8,12] and that these lock residues tend to be conserved [13]. Elsewhere we have shown that the closed loop folding hypothesis is consistent with the data derived from misincorporation proton-alkyl exchange experiments and from hydrogen exchange experiments [14] that have been used to derive foldons. Thus, closed loops may provide a preferable interpretation of this exchange data since they are contiguous, unlike the foldons, which may be disjoint [15-17].

To date, support for the closed loop hypothesis has largely been based on sequence analysis and equilibrium protein structures and has received less attention in mainstream protein folding studies. Here we challenge the hypothesis using results from *in vitro* protein folding experiments and from observations of dynamic protein structures. It is known from kinetic experiments that the folding rate of a protein correlates well with total contact distance (TCD) [18]; evaluating this metric across a subset of residues, including derived lock residues and their contacts, we note a marked improvement in this correlation, suggesting that that the lock residues and their neighbors together form the folding core of the protein. To address challenges in identifying lock residues, we consider two alternative approaches for identifying this core. Methods based upon the structural and chemical properties of residues, and upon high temperature molecular dynamics simulations each produce significant overlap with the sets of locks plus contacts. We next consider evidence for loop structures of length close to 25-30 residues. Coarse-grain elastic network studies show that protein residues with high force constants (and hence greater stability) tend to have a spacing of about 24 residues, compatible with the closed loop hypothesis. In earlier work, we found an association between sites of ligand binding and a measure of increased residue stability [19]; we here show that, across a set of diverse protein structures, loops of length 15-30 residues are more prevalent for cases in which at least one end of the loop is in a ligand binding site. Taken together, the results confirm that the residues predicted to mediate closed loop formation play an important role in protein folding, but it is also reasonable to conclude that the role of the lock residues has previously been over-emphasized as residues neighboring the lock residues are also important. The scope of the hypothesis and its relevance to protein stability and to drug design is discussed.

2. METHODS

2.1 Determination of closed loops. All loops of length 12-50 residues with minimum heavy-atom distance of 6 Å were determined. The contact region was scored according to the number of contact

neighbors, conservation, and hydrophobicity [13], evaluated over a window of 1, 3 or 5 adjacent residues. The highest scoring loop was determined first, and subsequent loops were identified such that there was minimum overlap between loops (but a given lock region frequently participated in two separate closed loops). The locks contained between 2 and 8 residues, with most having 4-6 residues (this is the full set of lock residues). For the TCD calculations, restricted to the 43 proteins where NMR structures were available, two further refinements were made. Firstly, lock residues that did not form persistent interactions across the ensemble of NMR structures were eliminated, generally leaving a set of 2-4 NMR-refined lock residues. Secondly, a minimal pair of two residues was selected to form the lock based on (a) good interactions, as observed using molecular graphics and (b) having a large number of neighbors (see the electronic supplementary material, table S1 and figure S1). The lock residues are found in all types of secondary structure, i.e. within helices, sheet, loops and β-turns and may overlap the junction between any two such secondary structures. Further details of the method for deriving loops are given in Chintapalli et al. [14].

2.2 Distribution of loop length. In order to analyze the distribution of chain return lengths in proteins, two hundred and seventy proteins, with lengths of at least 100 residues, and with well-characterized ligand-binding residues, as reported in the LPC database [20] were taken from the PDBSelect 25% list [21,22]. For each protein, all loops were found where the ends contacted to within 6 Å. The LPC database was used to identify loops for which at least one of the two loop end residues was contained within a ligand binding site.

2.3 Rigidity profiles. The rigidity profiles composed of residue-by-residue force constants were calculated from the conformational fluctuations observed for a set of 98 proteins taken from the set of Yang and Bahar [23]. Each force constant characterizes the difficulty of displacing the residue in question within the overall protein structure. The proteins were represented as coarse-grain elastic networks with 2-3 pseudo atoms per residue. We use an elastic network model where all the harmonic springs

connecting pseudo atoms less than 9 Å away have the same Hooke's Law force constant γ = 0.6 kcal mol^{-1} Å^{-2}. The elastic system is initially, by definition, in its equilibrium state and will undergo deformations around the equilibrium during the simulations because of the random displacement term in the Brownian dynamics equation of movement. In order to compare proteins of different sizes the force constants were re-expressed in units of standard deviation with respect to the mean for each protein (Z-scores) [24,25]. An autocorrelation was carried out of both the re-expressed force constants, k'_i, and of 99999 sets where the re-expressed force constants were randomized within the each protein. The significance of the peak in the autocorrelation was assessed by evaluating the ratio of the average value of $k'_i \times k'_j$ over the range 22 – 27 to that of the combined preceding range 15-20 and the following range 30-35. This was compared with that arising from the corresponding randomly generated values.

2.4 **Total contact distance**. The relationship between protein structure and folding rate has been well-established with the observation that the log of the folding rate, ln k_f, correlates with contact order, CO [26], which is defined as

$$CO = \frac{1}{L \times N} \sum_{N} |i - j| \qquad (1)$$

where N is the number of pairs of residues (i,j) that are in contact with one another, the metric $|i-j|$ describes the separation of the residues i and j in the chain, and L is the length of the protein. Alternatives to contact order, namely absolute contact order [27], long range order [28,29] and total contact distance, TCD, have been proposed. Here we have used TCD (equation 2), because TCD is insensitive as to whether immediate neighbors are included or not [18]. This makes TCD ideal for use in calculations based on selected subsets of residues and contacts. The use of CO or TCD with a subset of residues can also be justified by reference to work on the relationship between loop length and contact order [5].

$$TCD = \frac{1}{L^2} \sum_{N} |i - j| \qquad (2)$$

There are several sources of kinetic data for both two-state and multi-state proteins for use in such correlations [18,28,30-33], providing data for about 90 proteins. The kinetic data used in the correlation between the log folding rate and TCD was selected largely from Zou and Ozkan [33], for the 43 unique two-state proteins where NMR structures are available (NMR structures give some indication of the fluctuations present in the protein and hence the more important contacts within a lock). Two different sets of interactions were used in equation (2). These were (i) all residue-residue contacts, (ii) the interactions between the minimal pairs of lock residues plus interactions of the residues that contact the minimal pairs. (Other combinations are discussed in supplementary material).

In each case, a control was carried out by determining TCD for an equivalent number of contacts chosen by selecting the residues randomly; the significance of the real correlation was given by determining the proportion of cases that had a more extreme value of r^2. In common with other similar studies, we have primarily omitted 2 intervening residues in determining $|i - j|$ (i.e. we have excluded 1,2 and 1,3 contacts), but we have also studied the correlation between TCD and ln k_f for omission of between 3 and 80 residues.

2.5 Residue property-based approach to core identification

As an alternative approach to determining the protein core, the correlation between TCD and ln *kf* was determined following the removal of different percentages of the residues according to their connectivity or hydrophobicity values; residues were ranked according to their connectivity or hydrophobicity, and the residues with the lowest rank were removed first. The connectivity of a residue was defined as the number of other residues, at a distance of at least two residues in the chain, within a heavy atom distance of 6 Å. Hydrophobicity was calculated according to the octanol-water partition coefficient [34]. Where multiple residues had the same rank, random sets of residues fulfilling the criteria were removed, the mean correlation for 1000 repetitions of this process being output. In order to investigate whether clusters of moderately hydrophobic residues were more important than isolated highly hydrophobic residues,

residues were also removed according to the product of connectivity and hydrophobicity (where connectivity and hydrophobicity were scaled to between 0 and 1, with 1 representing the highest connectivity or hydrophobicity). We thus evaluated what percentage of residues could be eliminated while still retaining a good correlation between TCD and ln *kf* (a good correlation being similar to that of previous published work [18,26,28,35,36]. The optimal percentage of residues to be removed was determined by comparison with 1000 random removals at each percentage point, to generate statistics by a Monte Carlo approach. The calculations were carried out using Mathematica.

2.6 Molecular Dynamics Simulations

In order to investigate the participation of lock pairs and neighbors in initiating folding events, we analyzed molecular dynamics (MD) refolding simulations of four of the proteins, as described in electronic supplementary material.

3 RESULTS

3.1. Rigidity profile and the distribution of loop lengths. Evidence was found to support a preference for loop lengths in protein structures of close to 25 amino acids. We examined the distribution of chain return lengths from a diverse set of 270 proteins taken from PDBSelect25 [21,22]. Considering all such chain returns, our results reproduce those of Berezovsky et al. [1] (dashed red line, figure 1A). However, where one end of the loop is part of a ligand binding site, the distribution of loop lengths shows a more marked peak (solid black line, figure 1A), with the peak of this broad distribution corresponding to loops of length ~26 amino acids. The relevance of this to protein folding lies in the observation that residues in the folding nucleus tend to have a high number of residue-residue contacts [37,38], while recent work has shown that this property is shared by residues in ligand binding sites [19]. The common theme linking these two observations is entropy, since there is less loss of entropy on binding of a molecule to a rigid binding site [19]. It therefore appears that substrate binding and protein folding tend to use the same low

entropy regions, a principle that can be exploited in drug design by tailoring drugs to bind to lock residues and the associated core residues.

If the lock residues were part of the folding nucleus, they would be expected to have a greater number of residue-residue contacts, and to be held more rigidly within the protein structure. Examination of residues with a high degree of stability, as determined by coarse grain simulations, indicated a preference for mutual separation distances of around 24 residues. Thus, the rigidity profiles were determined using Brownian dynamics simulations in which each protein was modeled as an elastic network. A force constant was then assigned to each residue according to the magnitude of the fluctuations in the protein conformation; this force constant indicates the rigidity of a residue within the overall protein structure. The autocorrelation of residues with positive reduced force constants, k'_i are given in figure 1B (negative force constants were set to zero to avoid false positives). A clear broad peak was again observed for an inter-residue separation (i.e. loop length) of 22-27, giving similar results to those in figure 1A, and contributing to the evidence that loops of length ~25 residues are significant in protein structure. Monte Carlo analysis showed the peak to be highly significant ($p<10^{-5}$).

3.2 Relationship between total contact distance and folding rate. We have shown that the correlation between total contact distance (TCD) and the log folding rate of a set of proteins, ln k_f, was improved when the metric was applied to a reduced set of protein residues derived through application of the closed loop hypothesis. In the past, much interest has been focused upon identifying specific residues that play a key role in protein folding [37,38], and upon the prediction of folding rates from protein structure [18,26-28,36]. Thus, a number of experimental studies have noted a strong correlation between the log of the protein folding rate, ln k_f, and certain protein structure-derived metrics, namely length and structural class [30,39], number of contacts [40], contact order [26,27,36], absolute contact order (ACO = CO × L) [27,41], ACO with corrections [39], long range order [28] and TCD [18]. The correlation arises because both the number of contacts and the sequence distance per contact are important contributing factors to the kinetics of folding for two-state proteins, which, with some exceptions [42], have no intermediates

between the denatured state and the folding state. For two state proteins, while the folding rate correlates well with the topology (contact order), it correlates poorly with length, and so ACO does not work as well as CO, but for three state proteins or a mixture of peptides, two state and multistate proteins, ACO works better than CO [27]. The correlation between ln k_f and TCD for our set of 43 two-state proteins is shown in figure 2A. The correlation coefficient squared, r^2, is 0.65, a little smaller than the value of 0.77 reported by Zhou and Zhou for a similar analysis on a smaller set of 28 proteins [18]. Our 43 proteins were selected as two-state folders and include peptides; for this relatively diverse collection neither CO nor ACO works as well as TCD (supplementary figures 2A-3A). By analogy with ACO, we also tested ATCD (defined here as TCD × L) but this does not work as well as TCD (supplementary figure 4A).

When the evaluation of TCD was restricted to the interactions between the NMR-refined lock residues, r^2 decreased considerably; reducing each lock to a minimum pair of just two residues also gave insignificant results (not shown). In these simple applications, given in supplementary table 2, the closed loop hypothesis fails because the lock residues alone are insufficient to yield a good correlation.

An improved correlation was obtained by evaluating TCD across the lock residue minimum pairs and those residues in contact with them, with an r^2 of 0.76 ($p < 1\times10^{-4}$), (figure 2B). The protein locks-and-neighbors derived core identified by this result contained a mean of 101 contacts and 30 residues per protein, representing close to 38% of all contacts and 43% of all residues (see table 2 and supplementary figure S1). Thus, although this core represents a considerably reduced subset of the total residues, performing the analysis over this core gave a substantial improvement in the correlation between TCD and the log folding rate, suggesting that these residues play a key role in determining protein folding rates. (ATCD, ACO and particularly CO also gave improved correlations over this reduced subset, as shown in supplementary figures S2B-S4B). With just seven exceptions, the sets comprising lock residue pairs and their contacts included all of the originally identified NMR lock residues. The majority of contacts in the sets were in the same secondary structural elements as the lock residues.

Conversely, evaluation of TCD over interactions that did not include the NMR-refined lock residues or the lock pairs and their contacts resulted in insignificant correlations, as shown in supplementary table S2.

The failure to obtain a significant correlation in the absence of the lock residues indicates that these residues contribute to protein folding.

Our results in this section concur with those of Trifonov et al. regarding the prominence of loops of length ~25 amino acids [2]. However, it could be argued that this result merely reflects an artifact, such as the stiffness of the protein chain, rather than any property of locks related to protein folding. To address this issue, we monitored the variation of r^2 from figure 2A with loop length. In these calculations, the minimum loop length used in the TCD calculations varied from 2 to 54 intervening residues. The results in figure S5-S6 show that the correlation coefficient varies more closely with the distribution of the lengths of closed loops than with the distribution of loops in general. (Similar results, shown in supplementary figure S7, show that loops of length 25-40 are absolutely essential for a good correlation between TCD and ln k_f; but they also show that loops longer than ~45 residues make a significant contribution to the correlation).

We note that, although clear criteria have been determined for identifying lock residues, and hence the core, there are nevertheless some subjective decisions involved in determining and applying these criteria. We therefore consider alternative methods for finding the core, involving structural and chemical properties of residues.

3.3 Identifying the core from structural and chemical properties of residues. An approach based upon residue hydrophobicity identified protein cores which significantly overlapped the lock-pair plus contact sets. The correlation between TCD and ln k_f resulting from the removal of a given percentage of residues from each protein using a given measure (hydrophobicity and or connectivity) was calculated, and compared to statistics of the correlations arising from the random removal of the same number of residues. The removal of between 53% and 65% of residues by hydrophobicity resulted in a correlation superior to that calculated for 95% of the random sets. The first point gave a value of r^2=0.54 while the latter gave a value of r^2 of 0.46; the latter point (figure 3) was chosen, resulting in a small core of residues. Neither removal of residues according to connectivity nor the product of connectivity and

hydrophobicity yielded significant correlations (figure 3); this is in line with ideas on hydrophobic collapse [7] and down-plays the importance of highly connected nodes in folded proteins [37]. As in the TCD v ln k_f correlations (figure 2B, table 1, supplementary table S2), we required the presence of a larger set of residues that was generally about four times larger than the set of lock residues.

The mean size of the core that remains from the hydrophobicity method (referred to as the TCD-based core, figure 3) is, at 35% of the total number of residues (supplementary table S3), similar to the mean proportion of locks and neighbors, of 42±11%. Here, however, the core was constrained to be a similar proportion in each protein. The residues common to the TCD-based core and the locks and neighbors of supplementary table S1 together comprise about 58±11% of the residues and are recorded in supplementary table S3. A comparison of the two alternative cores is difficult because they were generated using different criteria, giving rise to different sizes: the TCD core was constrained to contain 35% of the residues and was based primarily on hydrophobicity while the locks and neighbors core varied between 24 and 65% and was determined by considering other factors in addition to hydrophobicity. Nevertheless, we assessed whether the TCD-based cores included pairs of residues for each lock (but not necessarily all the lock residues) and whether the TCD based cores reproduced more of the lock residues than would be expected at random. The only proteins that did not satisfy either of these criteria were 1gab, 2hqi and 2jwt and for these proteins TCD-derived core neighbors could play a similar role). For some proteins, *e.g. 1w4j* and *1ryk* the lock residues disappeared at 57%, 62% and 64% removal respectively, i.e. just below the 65% threshold (details are given in supplementary table S3).

3.3.1. MD refolding simulations

The simulations provided no evidence that nucleation involved a small set of residues (c.f the number of lock residues) but rather that nucleation involved a larger number of residues (c.f. the number of locks and neighbors). Moreover, the lock residues and their neighbors were prominent in this nucleation. For each pair of residues, the fraction of snapshots from the simulations in which the pair were in contact was calculated. This was compared to the same statistic averaged over all pairs at the

same contact distance; a resulting log ratio identified pairs that were statistically more likely to form contacts than would be expected from their distance apart in the protein chain.

In acyl-coenzyme A (PDB code 1NTI), around half of the 73 contacts identified as being more likely to form contacts in the simulations (log likelihood < -0.3) were contacts between residues also observed in both the lock pairs plus neighbors and in the TCD-derived core, as indicated in supplementary figures S8-S10. Given that the lock pairs plus neighbors represent only 38% of contacts in the original protein, this is a significant result ($p<0.05$, supplementary table S3). Similar results were identified for proteins G and L (supplementary table S3 and figures S9, S10). We would not expect exact agreement between the different cores as the fine details of the folding pathway may be force field dependent [43]. In addition, p-values calculated for the reproduction of the lock residues are less than 0.05 for each of the three proteins.

DISCUSSION

4.1. Evidence for 25mers

The loop length data and the data on the spacing of high force constant residues (figure 1) ties the observation of closed loops of around 25 residues more closely with protein folding since it associates the ends of the ~25mer loops with regions of high connectivity and or rigidity, which are themselves linked with protein folding [37,38]. The study on the variation of the TCD v ln k_f correlation with loop length (supplementary figure S7) indicates that the factor of 25 is not merely an artifact of protein structure and of peptide persistence length. Additionally, there is an interesting link between protein folding and ligand binding implicit in figure 1A. The link is implicit via the involvement of connectivity, since the folding nucleus and ligand binding sites are associated with regions of high connectivity [19,37]. Thus, because ligands are able to bind to regions involved in stabilizing the fold, they may also in some cases assist with fold stabilization. This has been seen very powerfully in the use of tightly binding ligands to stabilize

flexible structures such as G-protein coupled receptors (GPCR), and hence facilitate crystallization [44]. Similar principles may underlie the mechanism of pharmacological chaperones (small molecules that assist with the folding of proteins), as in the binding of SR49059 to the vasopressin receptor [45]. In both of these GPCR-based examples, the lock region and the ligand binding region occur within the same region of the transmembrane helical bundle. Evidence for closed loops of around 25 residues is also implicit in the TCD studies (figure 2B) since the mean length of the closed loops in this small sample of 43 proteins is 27 residues.

Since the lock residues and their neighbors are predicted to play some role in protein folding, these lock residues are given in table 1 for acyl-coenzyme A binding protein (pdb code 1NTI), indicating that this approach is able to generate useful molecular level information that is relevant to the folding process.

4.2 TCD correlations

The significance of the strong correlations between TCD and the log of the experimental folding rate in figure 2B is twofold. Firstly, the correlation in figure 2B involving subsets of residues predicted to form locks gives as strong a correlation as those previously observed [18,26-28,36] even though these new correlation results are based on ~60% fewer residues. The improved correlation over a reduced set of residues supports the idea that protein folding is driven by a subset of residues. Secondly, because the key residues were identified through application of the closed loop hypothesis, this suggests that the hypothesis may provide a valuable paradigm for understanding protein folding. Some support for the role of a subset of residues centered on the locks also comes from the TCD-based correlations/eliminations (figure 3) (and from the MD simulations, supplementary figures S8-S10, since these distributions were shown to overlap with the lock pairs and neighbors, supplementary table S3). The different percentages of residues used in each method and the need to randomly eliminate residues with equal rank in the TCD-based approach may have contributed to some of the differences observed between the two alternative cores (cf figures 2B, 3). In addition, the MD-based core (supplementary figures S8-S10) will to some

extent be dependent upon the force field, even though the overall picture to emerge from the MD simulations should be reliable [43].

Some additional support for expanding the closed loop hypothesis to include neighboring residues also comes from Φ-value analysis, as discussed in Supporting Information. In contrast, the closed loop hypothesis places much emphasis on the lock residues: the weight of evidence suggests that this is somewhat simplistic.

4.3. The relative importance of non-lock residues

Although we set out to investigate the role of lock residues, it is very clear from the techniques applied here that knowledge of the lock residues alone does not provide a clear and sufficient description of protein folding. Several of the techniques used, as exemplified by the data in figure 2B (TCD v ln k_f correlations), figure 3 (TCD v ln k_f correlations / eliminations), supplementary figures S8-S10 (MD simulations) and supplementary figure S11 (Φ-value analysis) strongly implicate a much larger protein core that is four or five times larger than the set of lock residues, but that is nevertheless much smaller than the set of all residues. This protein core has similarities to the concept of the extended nucleus described by Fersht [5]. The overlap between the locks and neighbors, the TCD-based core (figure 3) and the core derived from the MD simulations (supplementary figures S8-S10) indicates a common set of important residues that includes the lock residues.

Thus from this work, and previous studies [5], it seems that the lock residues are not the only residues with near-native structure in the transition structure for folding or comprising the protein core. For example, in the WW domain (code 1E0L), the largest Φ-values are for residues in β-turns (residues 14, 15 and 26), which are not part of the locks. These two β-turns are strategically placed to facilitate the folding of loops 8-22 and 20-36, which we have identified as closed loops. Thus in 1E0L, the β-turn may be more sensitive to substitution than the locks with regard to the formation of the closed loop, especially as the loop closure is not driven by a single residue. However, in the MD simulations on 1E0L, β-turns

residues 14, 15 and 16 make a few short range contacts, while lock residues 8, 20, 22 and 36 make significant long-range contacts, results not shown. A similar effect is observed in protein G (pdb code 3GB1) [46], protein L (pdb code 2PTL) [47] and phosphotransferase (pdb code 1FYN) [48]. Thus, despite a slight preference for lock residues to have high median Φ-values (supplementary table S6), a high Φ-value does not necessarily imply a key role in nucleation [48,49]. In other proteins, high Φ-values are recorded for residues in the vicinity of the locks, particularly those in the same secondary structural element as the lock, indicating that for the lock to form, the secondary structure in which it resides may also have to form [6,50,51]. A similar conclusion was drawn from a re-evaluation of the native-state hydrogen exchange experiments in the light of the closed loop hypothesis[14].

Rustad and Ghosh [39] found that absolute contact order could be significantly improved by explicit corrections for nested (this could include omega loops [52]) or linked (i.e. overlapping) loops. While such loops do not feature significantly in the original closed loop hypothesis, as formulated by Berezovsky et al., as the closed loops should be non-overlapping (bar a small overlap of about 5 residues), we find that a reasonable number of nested and linked loops are taken into account in our approach via the lock-neighbors, as illustrated in supplementary figure S12 for protein 2rpn, a yeast SH3 domain; this may be one reason why it is necessary to supplement the locks with their neighbors.

In a similar vein, local contacts have no formal place within the closed look hypothesis, and indeed non-local contacts are known to dominate the barrier-crossing process [33]. Local contacts nevertheless play a role in folding [33,53] and supplementary figure S7 shows that 'local TCD' correlates with ln k_f (a similar correlation with 'local contact order was shown by Zou and Ozkan [33]). Indeed local contacts have been discussed above with regard to high Φ-values and native-state hydrogen exchange experiments. The inclusion of lock neighbors automatically introduces a number of local contacts into TCD or related metrics, at least in the vicinity of the locks, as shown by the neighboring circles in supplementary figure S12. Thus our modification of the closed loop hypothesis includes local

interactions, overlapping loops and nested loops that interact with the lock residues, possibly giving increased prominence to the lock residues.

Elsewhere we have observed that it is not yet possible to determine the lock residues precisely, since different authors using slightly different methods may only agree to within one or two residues [14]. Supplementary table S6 indicates that Φ-values may offer an indication as to whether residues participate in lock regions but the data is far from definitive. Thus, although we have identified a significant set of residues, the precise functional distinction between lock residues and non-lock residues is not clear at the present time. Based on the current TCD calculations, for some dynamic applications it may therefore be more useful to distinguish between core (c.f. extended nucleus) and non-core residues. Here, the core could be derived from the lock pairs and neighbors [14], from the elimination of residues during TCD-based simulations (figure 3) or from MD simulations. For some proteins such as 2jwt, the locks plus neighbors core is a relatively low percentage (24%) of the total number of residues (13 residues out of a total of 54), while for others (e.g. 1wiu), it is a relatively high percentage (65%), as shown in table 2.

There are long range contributions to folding. Supplementary figure S7 shows that loops as long as 80 residues make a significant contribution to the correlation shown in figure 2A. This is consistent with the observations that locks from different closed loops tend to cluster together [14]. Thus, while the closed loop hypothesis is important for determining the lock residues, the folding process is certainly not local to the interactions within the closed loop.

4.4. Protein stability

Since there is much interest in modifying protein stability, e.g. as an aid to crystallization [44,54], the concept of identifying the protein core from the closed loops could be very useful, (a) for ensuring that the core is maintained under mutagenesis and (b) for identifying areas of low stability (e.g. non-core regions) where an increase in stability could be most beneficial. Increases in stability can be engineered through mutation [54-56] or can come from a molecular chaperone binding to the lock residues / protein core [45]. The protein cores for acyl-coenzyme A binding protein and the Ras-binding domain of c-Raf-1

(PDB codes 1NTI and 1RFA respectively) are shown in figure 4; the cores for the 43 proteins are shown in supplementary figure S1.

CONCLUSIONS

We find additional new evidence for the importance of ~25mer closed loops from both the autocorrelation of residues with high force constants, as determined by the elastic networks, and from the distribution of loop lengths where one end of the loop is part of the ligand binding site. However, we find that the closed loop hypothesis is somewhat lacking in that the locks themselves are certainly not sufficient for obtaining a good correlation, but rather that the locks need to be supplemented by considerably more residues. This additional requirement is not evident from analysis of structure or sequence alone but arises when the protein dynamics is considered. However, the closed loop hypothesis may nevertheless be a useful tool for guiding experiments to determine the nature of this core, which surrounds the lock residues.

ACKNOWLEDGEMENTS

We acknowledge support from the University of Essex (SVC) and the Royal Society for a Theo Murphy Blue Skies Award (CAR). The molecular dynamics simulation data was originally generated on Blue Gene for an alternative project and was kindly provided by Dr Jed Pitera, IBM Almaden Research Center, San Jose, CA, USA.

Figure Legends

Figure 1. Further evidence for loops of length ~25 residues. (A) The distribution of loop lengths in protein structures in general (dashed line) and the distribution of loop lengths where one end of the loop is

in a ligand binding site (solid line). The notable feature of the graph is the marked increase in the height of the peak at 26 (solid line), corresponding to the proposed mean length of the closed loops. The analysis is over 250 proteins from the PDBSelect25 set. Ligand binding sites were identified from the LPC database. (B) Autocorrelation, C, for reduced force constants, k', greater than zero, is defined as $\Sigma k' \times k'_L$, where k'_L represents the value of k' L residues further along the sequence (C is normalized according to the number of residues at distance L apart in each protein). The notable feature is the marked increase in the height of the peak at a residue separation of 24, corresponding to the proposed mean length of the closed loops.

Figure 2. The relationship between the log of the folding rate, ln k_f, and total contact distance for a set of 43 two-state proteins (A) evaluated over all residues and (B) evaluated over the minimal pair of lock residues plus residues that contact them. For (A), the outliers (for no obvious reason) are 1BA5, 1K8O, 1PSE, 1YZA, 1N88, 1PKS, 2AX5; for (B) the outlier is 1K8O.

Figure 3. Variation of the correlation coefficient, r^2, between TCD and ln k_f as the percentage of residues progressively removed increases. The red, blue and black solid lines indicate the values of r^2 when residues are removed according to their octanol partition coefficient, connectivity and the product of the two respectively. Black dotted lines at increasing vertical values indicate 1%, 5%, 50%, 95% and 99% percentile r^2 values respectively, obtained by random removal of given percentages of residues from each of the set of proteins. Thus the 95% significance region lies above the light grey area. The point corresponding to figure 2B is shown as a green circle. The red octanol line almost reaches 99% significance at 61% removal.

Figure 4. The protein cores for (A) 1nti and (B) 1rfa, as determined from the minimal pairs of lock residues (opaque, spacefill) and their equally important contacts (transparent, spacefill). The first closed loop is colored red, the second green.

References

1. Berezovsky, I.N., Grosberg, A.Y., & Trifonov, E.N. 2000 Closed loops of nearly standard size: common basic element of protein structure. *Febs Letters* **466**, 283-286.
2. Trifonov, E.N. & Berezovsky, I.N. 2003 Evolutionary aspects of protein structure and folding. *Curr. Opin. Struct. Biol.* **13**, 110-114.
3. Ittah, V. & Haas, E. 1995 Nonlocal interactions stabilize long range loops in the initial folding intermediates of reduced bovine pancreatic trypsin inhibitor. *Biochemistry* **34**, 4493-4506.
4. Fersht, A.R. & Daggett, V. 2002 Protein folding and unfolding at atomic resolution. *Cell* **108**, 573-582.
5. Fersht, A.R. 2000 Transition-state structure as a unifying basis in protein-folding mechanisms: contact order, chain topology, stability, and the extended nucleus mechanism. *Proc Natl. Acad. Sci. U. S. A* **97**, 1525-1529.
6. Gianni, S., Guydosh, N.R., Khan, F., Caldas, T.D., Mayor, U., White, G.W., DeMarco, M.L., Daggett, V., & Fersht, A.R. 2003 Unifying features in protein-folding mechanisms. *Proc Natl. Acad. Sci. U. S. A* **100**, 13286-13291.
7. Dill, K.A., Ozkan, S.B., Shell, M.S., & Weikl, T.R. 2008 The protein folding problem. *Annu. Rev. Biophys.* **37**, 289-316.
8. Berezovsky, I.N., Kirzhner, V.M., Kirzhner, A., & Trifonov, E.N. 2001 Protein folding: Looping from hydrophobic nuclei. *Proteins* **45**, 346-350.

9. Aharonovsky, E. & Trifonov, E.N. 2005 Protein sequence modules. *J. Biomol. Struct. Dyn.* **23**, 237-242.

10. Aharonovsky, E. & Trifonov, E.N. 2005 Sequence structure of van der Waals locks in proteins. *J. Biomol. Struct. Dyn.* **22**, 545-553.

11. Voigt, C.A., Martinez, C., Wang, Z.G., Mayo, S.L., & Arnold, F.H. 2002 Protein building blocks preserved by recombination. *Nat. Struct. Biol.* **9**, 553-558.

12. Lamarine, M., Mornon, J.P., Berezovsky, N., & Chomilier, J. 2001 Distribution of tightened end fragments of globular proteins statistically matches that of topohydrophobic positions: towards an efficient punctuation of protein folding? *Cell Mol. Life Sci.* **58**, 492-498.

13. Yew, B.K., Chintapalli, S.V., Upton, G.G., & Reynolds, C.A. 2007 Conservation of closed loops. *J Mol. Graph Model* **26**, 652-655.

14. Chintapalli, S.V., Yew, B.K., Illingworth, C.J., Upton, G.J., Reeves, P.J., Parkes, K.E., Snell, C.R., & Reynolds, C.A. 2010 Closed loop folding units from structural alignments: experimental foldons revisited. *J Comput. Chem* **31**, 2689-2701.

15. Bai, Y., Sosnick, T.R., Mayne, L., & Englander, S.W. 1995 Protein folding intermediates: native-state hydrogen exchange. *Science* **269**, 192-197.

16. Krishna, M.M., Lin, Y., Rumbley, J.N., & Englander, S.W. 2003 Cooperative omega loops in cytochrome c: role in folding and function. *J Mol. Biol.* **331**, 29-36.

17. Krishna, M.M., Maity, H., Rumbley, J.N., Lin, Y., & Englander, S.W. 2006 Order of steps in the cytochrome C folding pathway: evidence for a sequential stabilization mechanism. *J Mol. Biol.* **359**, 1410-1419.

18. Zhou, H. & Zhou, Y. 2002 Folding rate prediction using total contact distance. *Biophys. J.* **82**, 458-463.

19. Illingworth, C.J., Scott, P.D., Parkes, K.E., Snell, C.R., Campbell, M.P., & Reynolds, C.A. 2010 Connectivity and binding-site recognition: applications relevant to drug design. *J Comput. Chem* **31**, 2677-2688.

20. Sobolev, V., Sorokine, A., Prilusky, J., Abola, E.E., & Edelman, M. 1999 Automated analysis of interatomic contacts in proteins. *Bioinformatics*. **15**, 327-332.

21. Hobohm, U., Scharf, M., Schneider, R., & Sander, C. 1992 Selection of representative protein data sets. *Protein Sci.* **1**, 409-417.

22. Hobohm, U. & Sander, C. 1994 Enlarged representative set of protein structures. *Protein Sci.* **3**, 522-524.

23. Yang, L.W. & Bahar, I. 2005 Coupling between catalytic site and collective dynamics: a requirement for mechanochemical activity of enzymes. *Structure.* **13**, 893-904.

24. Sacquin-Mora, S. & Lavery, R. 2006 Investigating the local flexibility of functional residues in hemoproteins. *Biophys. J* **90**, 2706-2717.

25. Sacquin-Mora, S., Laforet, E., & Lavery, R. 2007 Locating the active sites of enzymes using mechanical properties. *Proteins* **67**, 350-359.

26. Plaxco, K.W., Simons, K.T., & Baker, D. 1998 Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**, 985-994.

27. Ivankov, D.N., Garbuzynskiy, S.O., Alm, E., Plaxco, K.W., Baker, D., & Finkelstein, A.V. 2003 Contact order revisited: influence of protein size on the folding rate. *Protein Sci.* **12**, 2057-2062.

28. Gromiha, M.M. & Selvaraj, S. 2001 Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction. *J Mol. Biol.* **310**, 27-32.

29. Harihar, B. & Selvaraj, S. 2009 Refinement of the long-range order parameter in predicting folding rates of two-state proteins. *Biopolymers* **91**, 928-935.

30. DeSancho D. & Munoz, V. 2011 Integrated prediction of protein folding and unfolding rates from only size and structural class. *Phys Chem Chem Phys* **13**, 17030-17043.

31. Ouyang, Z. & Liang, J. 2008 Predicting protein folding rates from geometric contact and amino acid sequence. *Protein Sci.* **17**, 1256-1263.

32. Maxwell, K.L., Wildes, D., Zarrine-Afsar, A., De Los Rios, M.A., Brown, A.G., Friel, C.T., Hedberg, L., Horng, J.C., Bona, D., Miller, E.J. *et al.* 2005 Protein folding: defining a "standard" set of experimental conditions and a preliminary kinetic data set of two-state proteins. *Protein Sci.* **14**, 602-616.

33. Zou, T. & Ozkan, S.B. 2011 Local and non-local native topologies reveal the underlying folding landscape of proteins. *Phys Biol* **8**, 066011.

34. White, S.H. & Wimley, W.C. 1999 Membrane protein folding and stability: physical principles. *Annu. Rev. Biophys. Biomol. Struct.* **28**, 319-365.

35. Galzitskaya, O.V., Garbuzynskiy, S.O., Ivankov, D.N., & Finkelstein, A.V. 2003 Chain length is the main determinant of the folding rate for proteins with three-state folding kinetics. *Proteins* **51**, 162-166.

36. Paci, E., Lindorff-Larsen, K., Dobson, C.M., Karplus, M., & Vendruscolo, M. 2005 Transition state contact orders correlate with protein folding rates. *J Mol. Biol.* **352**, 495-500.

37. Vendruscolo, M., Dokholyan, N.V., Paci, E., & Karplus, M. 2002 Small-world view of the amino acids that play a key role in protein folding. *Phys. Rev. E.* **65**, 061910.

38. Vendruscolo, M., Paci, E., Dobson, C.M., & Karplus, M. 2001 Three key residues form a critical contact network in a protein folding transition state. *Nature* **409**, 641-645.

39. Rustad, M. & Ghosh, K. 2012 Why and how does native topology dictate the folding speed of a protein? *J Chem Phys* **137**, 205104.

40. Makarov, D.E., Keller, C.A., Plaxco, K.W., & Metiu, H. 2002 How the folding rate constant of simple, single-domain proteins depends on the number of native contacts. *Proc Natl. Acad. Sci. U. S. A* **99**, 3535-3539.

41. Plaxco, K.W., Simons, K.T., Ruczinski, I., & Baker, D. 2000 Topology, stability, sequence, and length: defining the determinants of two-state protein folding kinetics. *Biochemistry* **39**, 11177-11183.

42. Northey, J.G., Di Nardo, A.A., & Davidson, A.R. 2002 Hydrophobic core packing in the SH3 domain folding transition state. *Nat. Struct. Biol* **9**, 126-130.

43. Piana, S., Lindorff-Larsen, K., & Shaw, D.E. 2011 How robust are protein folding simulations with respect to force field parameterization? *Biophys. J* **100**, L47-L49.

44. Warne, T., Serrano-Vega, M.J., Baker, J.G., Moukhametzianov, R., Edwards, P.C., Henderson, R., Leslie, A.G., Tate, C.G., & Schertler, G.F. 2008 Structure of a beta1-adrenergic G-protein-coupled receptor. *Nature* **454**, 486-491.

45. Kobayashi, H., Ogawa, K., Yao, R., Lichtarge, O., & Bouvier, M. 2009 Functional rescue of beta-adrenoceptor dimerization and trafficking by pharmacological chaperones. *Traffic.* **10**, 1019-1033.

46. McCallister, E.L., Alm, E., & Baker, D. 2000 Critical role of beta-hairpin formation in protein G folding. *Nat. Struct. Biol* **7**, 669-673.

47. Kim, D.E., Fisher, C., & Baker, D. 2000 A breakdown of symmetry in the folding transition state of protein L. *J Mol. Biol* **298**, 971-984.

48. Northey, J.G., Maxwell, K.L., & Davidson, A.R. 2002 Protein folding kinetics beyond the phi value: using multiple amino acid substitutions to investigate the structure of the SH3 domain folding transition state. *J Mol. Biol* **320**, 389-402.

49. Hubner, I.A., Shimada, J., & Shakhnovich, E.I. 2004 Commitment and nucleation in the protein G transition state. *J Mol. Biol* **336**, 745-761.

50. Itzhaki, L.S., Otzen, D.E., & Fersht, A.R. 1995 The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J Mol. Biol.* **254**, 260-288.

51. Otzen, D.E., Itzhaki, L.S., elMasry, N.F., Jackson, S.E., & Fersht, A.R. 1994 Structure of the transition state for the folding/unfolding of the barley chymotrypsin inhibitor 2 and its implications for mechanisms of protein folding. *Proc Natl. Acad. Sci. U. S. A* **91**, 10422-10425.

52. Leszczynski, J.F. & Rose, G.D. 1986 Loops in globular proteins: a novel category of secondary structure. *Science* **234**, 849-855.

53. Ghosh, K. & Dill, K.A. 2009 Theory for protein folding cooperativity: helix bundles. *J Am. Chem Soc* **131**, 2306-2312.

54. Shibata, Y., White, J.F., Serrano-Vega, M.J., Magnani, F., Aloia, A.L., Grisshammer, R., & Tate, C.G. 2009 Thermostabilization of the neurotensin receptor NTS1. *J. Mol. Biol.* **390**, 262-277.

55. Serrano-Vega, M.J., Magnani, F., Shibata, Y., & Tate, C.G. 2008 Conformational thermostabilization of the beta1-adrenergic receptor in a detergent-resistant form. *Proc. Natl. Acad. Sci U. S. A* **105**, 877-882.

56. Schlinkmann, K.M., Honegger, A., Tureci, E., Robison, K.E., Lipovsek, D., & Pluckthun, A. 2012 Critical features for biosynthesis, stability, and functionality of a G protein-coupled receptor uncovered by all-versus-all mutations. *Proc Natl. Acad. Sci. U. S. A* **109**, 9810-9815.

Table 1. Molecular contacts (i.e. lock residues) for acyl-coenzyme A binding protein (pdb codes 1NTI/2ABD).

The Φ-values are given where these are available.

1NTI	Full	NMR-refined	Highest Φ-values
Loop 1	5	5	5(0.74)
	30,31,32	30,31	32(0.96)
Loop 2	28,29,30	30	-
	73	73	73 (0.7)

Table 2. The extent of the protein core as defined by residues neighboring the lock residues. The Table shows both the number of contacts and the number of residues involved in the correlations given in Fig. 1 (cf supplementary figure S1).

	Full set of contacts (cf Fig. 2A)		Lock pairs +contacts (cf Fig. 2B)		
Protein	Residues	Contacts	Residues	Contacts	ln k_f
1aey	58	221	31	110	2.09
1aps	98	426	53	213	-1.48
1ba5	53	171	22	65	5.91
1cis	66	272	33	128	3.87
1e0l	37	94	11	17	10.37
1e0m	37	101	13	29	8.85
1fex	59	207	22	58	8.19
1fkr	107	424	51	186	1.45
1g6p	66	260	21	59	6.30
1gab	53	189	17	44	12.7
1hdn	85	369	44	153	2.70
1idz	54	173	15	39	8.73
1imp	86	343	39	139	7.31
1k0s	151	613	63	250	7.44
1k8o	87	324	41	125	-0.71
1k9q	40	114	12	28	8.37
1l2y	20	42	6	7	12.4
1n88	96	398	51	219	2.02
1nti	86	364	35	139	6.96
1nyg	58	212	27	83	4.54
1o6x	81	271	22	54	6.63
1pba	81	293	23	65	6.80
1pks	76	325	46	191	-1.05
1pse	69	233	32	97	1.17
1rfa	78	307	27	85	8.36
1ryk	69	280	17	43	9.08
1srm	56	192	28	79	4.04
1ss1	60	234	21	65	11.48
1w4e	45	159	18	56	10.22
1w4j	51	170	18	53	12.25
1wiu	93	408	60	244	0.41
1yza	106	352	28	76	8.40
2ait	74	306	34	125	4.20
2ax5	99	350	51	157	2.58
2bth	45	146	15	39	11.78
2hqi	72	342	39	144	0.18
2jwt	54	191	13	16	10.53
2pdd	43	135	14	31	9.80
2ptl	62	245	36	108	4.10
2rpn	58	250	25	88	2.46
2vil	126	559	66	293	6.80
3gb1	56	201	23	72	6.30
3mef	69	242	24	73	5.30

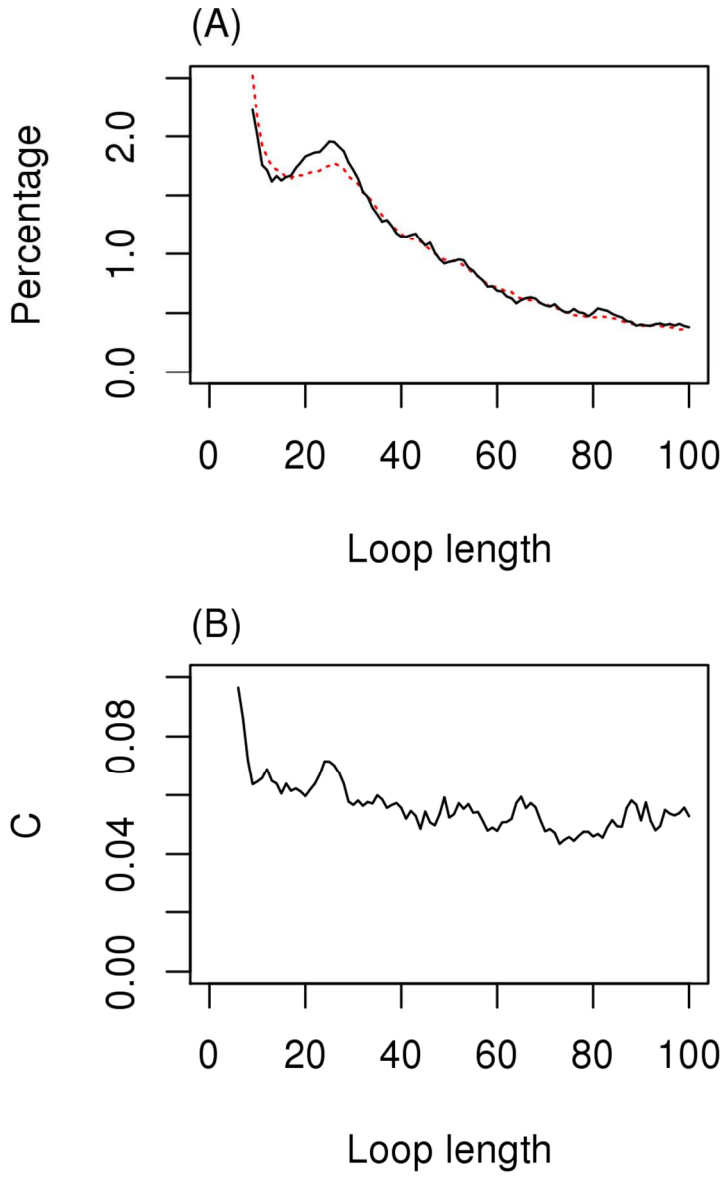

Figure 1. Further evidence for loops of length ~25 residues. (A) The distribution of loop lengths in protein structures in general (dashed line) and the distribution of loop lengths where one end of the loop is in a ligand binding site (solid line). The notable feature of the graph is the marked increase in the height of the peak at 26 (solid line), corresponding to the proposed mean length of the closed loops. The analysis is over 250 proteins from the PDBSelect25 set. Ligand binding sites were identified from the LPC database. (B) Autocorrelation, C, for reduced force constants, k', greater than zero, is defined as $\sum k' \times k'_L$, where k'_L represents the value of k' L residues further along the sequence (C is normalized according to the number of residues at distance L apart in each protein). The notable feature is the marked increase in the height of the peak at a residue separation of 24, corresponding to the proposed mean length of the closed loops.
81x129mm (300 x 300 DPI)

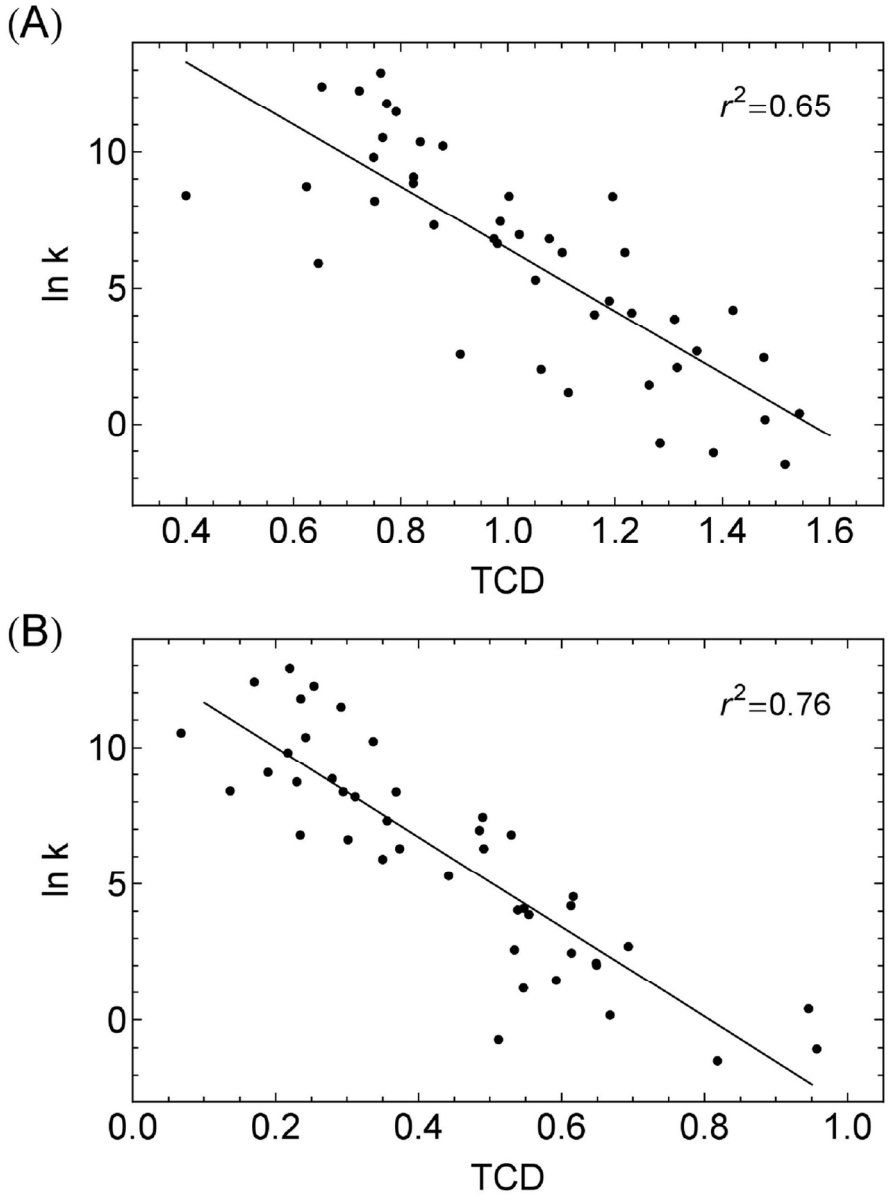

Figure 2. The relationship between the log of the folding rate, ln k_f, and total contact distance for a set of 43 two-state proteins (A) evaluated over all residues and (B) evaluated over the minimal pair of lock residues plus residues that contact them.
111x147mm (300 x 300 DPI)

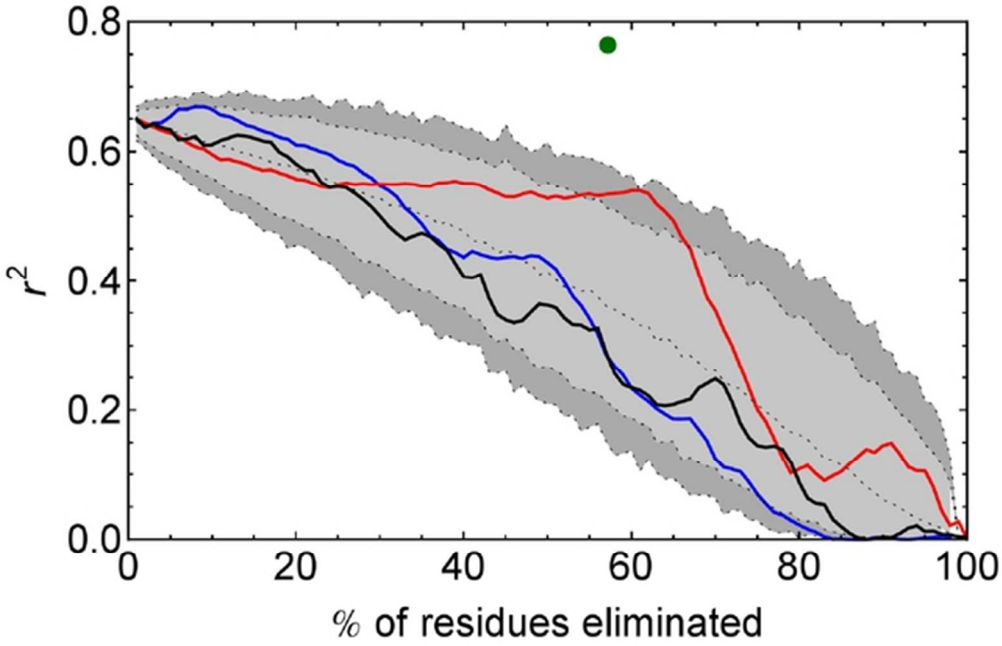

Figure 3. Variation of the correlation coefficient, r^2, between TCD and ln k_f as the percentage of residues progressively removed increases. The red, blue and black solid lines indicate the values of r^2 when residues are removed according to their octanol partition coefficient, connectivity and the product of the two respectively. Black dotted lines at increasing vertical values indicate 1%, 5%, 50%, 95% and 99% percentile r^2 values respectively, obtained by random removal of given percentages of residues from each of the set of proteins. Thus the 95% significance region lies above the light grey area. The point corresponding to figure 2B is shown as a green circle. The red octanol line almost reaches 99% significance at 61% removal.
54x34mm (300 x 300 DPI)

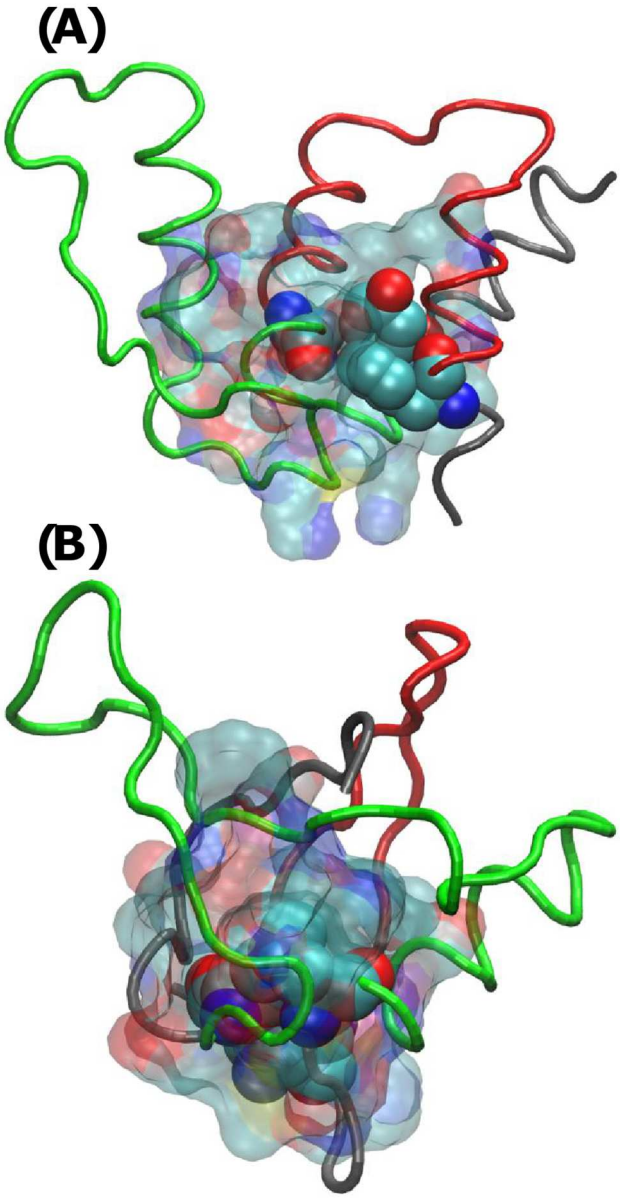

Figure 4. The protein cores for (A) 1nti and (B) 1rfa, as determined from the minimal pairs of lock residues (opaque, spacefill) and their equally important contacts (transparent, spacefill). The first closed loop is colored red, the second green.
82x159mm (300 x 300 DPI)