doi:10.1016/j.jmb.2006.06.061 J. Mol. Biol. (2006) 362, 151–171

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

# Massive Sequence Perturbation of the Raf *ras* Binding Domain Reveals Relationships between Sequence Conservation, Secondary Structure Propensity, Hydrophobic Core Organization and Stability

**F.-X. Campbell-Valois[1,2], Kirill Tarassov[1] and S. W. Michnick[1]***

[1]*Département de Biochimie Université de Montréal C.P. 6128, Succ. centre-ville Montréal, Québec Canada H3C 3J7*

[2]*Programme de Biologie Moléculaire, Université de Montréal, C.P. 6128 Succ. centre-ville, Montréal Québec, Canada H3C 3J7*

**Corresponding author*

The contributions of specific residues to the delicate balance between function, stability and folding rates could be determined, in part comparing the sequences of structures having identical folds, but insignificant sequence homology. Recently, we have devised an experimental strategy to thoroughly explore residue substitutions consistent with a specific class of structure. Using this approach, the amino acids tolerated at virtually all residues of the c-Raf/Raf1 *ras* binding domain (Raf RBD), an exemplar of the common β-grasp ubiquitin-like topology, were obtained and used to define the sequence determinants of this fold. Herein, we present analyses suggesting that more subtle sequence selection pressure, including propensity for secondary structure, the hydrophobic core organization and charge distribution are imposed on the Raf RBD sequence. Secondly, using the Gibbs free energies ($\Delta G_{F\text{-}U}$) obtained for 51 mutants of Raf RBD, we demonstrate a strong correlation between amino acid conservation and the destabilization induced by truncating mutants. In addition, four mutants are shown to significantly stabilize Raf RBD native structure. Two of these mutations, including the well-studied R89L, are known to severely compromise binding affinity for *ras*. Another stabilized mutant consisted of a deletion of amino acid residues E104–K106. This deletion naturally occurs in the homologues a-Raf and b-Raf and could indicate functional divergence. Finally, the combination of mutations affecting five of 78 residues of Raf RBD results in stabilization of the structure by approximately 12 kJ mol$^{-1}$ ($\Delta G_{F\text{-}U}$ is −22 and −34 kJ mol$^{-1}$ for wt and mutant, respectively). The sequence perturbation approach combined with sequence/structure analysis of the ubiquitin-like fold provide a basis for the identification of sequence-specific requirements for function, stability and folding rate of the Raf RBD and structural analogues, highlighting the utility of conservation profiles as predictive tools of structural organization.



*Keywords:* Raf *ras* binding domain; β-grasp ubiquitin-like topology; sequence entropy; core volume; secondary structure propensity

Abbreviations used: RBD, *ras* binding domain; PFAM, protein families; SMART, simple modular architecture research tool; SCOP, structural classification of proteins; CATH class, architecture, topology and homologous superfamily; FSSP, families of structurally similar proteins; MAPK, mitogen activated protein kinase; wt, wild-type; SH3, src-homology 3; Gdm-HCl, guanidinium-hydrochloride; TS, transition state.

E-mail address of the corresponding author: stephen.michnick@umontreal.ca

## Introduction

It has been known since the determination of the earliest protein structures that diverging polypeptide chains can adopt the same overall arrangement of their backbone atoms (e.g. topology).[1–3] While this could seem at odds with Anfinsen's principle that the structural information is encoded fully and completely in the sequence, it is rather an indication of the degenerate nature of the chemical information encrypted in polypeptide sequences. To uncover this code, the most instructive structural comparisons are based on proteins displaying no apparent evolutionary links, specifically in the absence of significant sequence homology and common biological functions, because it could then be reasonably argued that the few commonly constrained positions in the sequences are important for defining the structures. The comparisons of the primary structure of proteins adopting similar topology and of the biophysical characteristics of their folding and stabilization have been used to try to decipher the redundant messages embedded in amino acid sequences.[4–16]

In general, the analysis of sequence alignment of proteins adopting similar fold but low sequence identity highlights primarily conservation of hydrophobic core positions,[4–7,17] but yields little information about the precise role and interplay between the residues in stabilization and formation of the structure. The comparison of the folding reaction of several homologous and more distantly related proteins has indicated that their folding mechanisms are often comparable,[9,10,13,14,16] but in some cases significantly different[11,12] (these examples and others reviewed by Zarrine-Afsar *et al.*[18]). These results suggest that the folding mechanism may be encoded in the polypeptide sequence in a less constrained manner than originally thought. This hypothesis could provide an explanation for the diversity of structural forms in natural proteins, by which the most versatile topology would be favored. Surprisingly, few studies have combined thermodynamic studies and sequence alignment analysis in an integrated manner.[7] This type of approach holds the promise of novel insights about the relationships between sequence conservation, folding mechanism, stability and function. Databases combining sequence and structural information such as "protein families" (PFAM) or "simple modular architecture research tool" (SMART) and "structural classification of proteins" (SCOP), "class, architecture, topology and homologous superfamily" (CATH), "homology-derived secondary structure of proteins" (HSSP) or "families of structurally similar proteins" (FSSP) are useful tools to address these types of problems on structurally close or distantly related proteins, respectively.

The protein universe is not homogenous in the sense that some protein topologies have been selected and became more represented than others throughout evolution. For example, an estimated 80% of known structures adopt one of the 400 most frequent topologies of the 10,000 predicted to exist in nature.[19] This statement is confirmed by the low number of novel topologies discovered by structural genomics research programs that established experimental methodologies specifically designed to identify novel folds (consult http://www.jcsg.org/ and http://www.strgen.org/ for statistics).[20] A corollary of these observations is that most of known topologies are and therefore sequence information for these folds is rare. Rarely occurring folds pose a challenge to the study of structure stabilization and folding based on sequence constraints, because of the paucity of sequence information. The utilization of degenerated libraries to increase the sequence space covered by such poorly populated folds is a simple solution to this problem. Many approaches have been designed to introduce targeted randomization into the primary sequences of proteins and select mutants based on their capacity to fold. Woolfson and co-workers presented a strategy for selecting stable mutants based on resistance to chymotrypsin digestion.[21] Using this approach, they performed small scale co-variation of eight hydrophobic core residues in ubiquitin and reported equilibrium and kinetic properties of an over- and an under-packed mutant.[22,23] Due to the difficulty of establishing selection assays that rely solely on structural stability for every model protein of interest, most studies use specific functional assays to select for folding mutants.[24–28] The sequence perturbation strategies are particularly attractive, because they allow, without the unwanted bias embedded in natural sequence evolution, for expanding the sequence diversity available to specific protein topologies. Thus, the relationships between sequence, structure and function becomes practical to determine. The nature of evolutionary pressure imposed by structure conservation was partially revealed in such studies, yielding evidence that the folding rate is not optimized by evolution.[27,28] Building on these studies, we have recently reported a method to massively perturb the sequence of small proteins, and demonstrated its application to test the sequence variation tolerated at virtually every residue of c-Raf/Raf-1 *ras* binding domain (RBD).[15]

The most recognized biological function of the Ser/Thr kinase Raf is to activate the mitogen activated protein kinase (MAPK) pathway. The classical scheme of MAPK cascade activation is by recruitment of Raf to the membrane *via* binding its RBD to GTP loaded *ras*, which then leads to relief of the auto-inhibition of the kinase activity through phosphorylation at several sites on the Raf protein (reviewed by Wellbrock *et al.*[29]). At present, no structures of the heterodimeric complex between Raf RBD and *ras* have been reported. However, the complex between mutated Rap1A with charge reversal at position 31 to mimic *ras* and c-Raf/Raf-1 RBD has been used to model the Raf-*ras* complex.[30] Residues

located on the basic surface of Raf RBD were shown by mutagenesis to be essential for complex formation.[31] The Raf RBD is composed of 78 amino acid residues that form a globular structure, which is a member of the β-grasp ubiquitin-like (aka ubiquitin-roll) topology according to the SCOP database† (Figure 1(a)). This protein topology groups several superfamilies, some linked by putatively common evolutionary origins, while others appear to result from convergent evolution. Not surprisingly then, this topology represent one of the ten most highly occurring domains (e.g. superfold) in the protein universe with an estimated frequency of 1.1%.[19] The high occurrence of this topology yields a plethora of analogous structures that have very little sequence identity even within a superfamily. For example, the Raf RBD and ubiquitin, have approximately 12% residue identity based on alignment of sequences to secondary structure. Comparable sensitivity of the folding rate of these proteins to mutations and chemical perturbation has been demonstrated.[14] In the accompanying manuscript, we show by Φ-value analysis that the structure of the transition state (TS) of mammalian ubiquitin and Raf RBD share common characteristics.[16] Applied to the Raf RBD, the sequence perturbation strategy mentioned above, consisted of randomizing the polypeptide sequence within 13 discrete segments corresponding to secondary structure elements and selecting the variants able to fold based on their capacity to interact *in vivo* with *h-ras*.[15] In this study, the focus was on analyses of the tolerance to mutation of each position (e.g. sequence entropy) and the specific amino acid selection at each position (Figure 1). Specifically, we have discriminated between the functional and structural constraints at each conserved residue and shown that the conservation observed recapitulates the sequence variability observed in alignments of structural analogues recovered in SCOP β-grasp ubiquitin-like topologies. Herein, we discuss more subtle aspects of selection-pressures, including secondary structure propensity, hydrophobic core organization and charge distribution that are imposed on the Raf RBD sequence in the perturbation experiments and by natural evolution.

An important debate about sequence evolution concerns the specific conservation or not of residues displaying high Φ-values (those forming the folding nucleus).[32–35] On the other hand, computer simulations and experiments suggested that native state stability and function are the major determinants of sequence conservation in the SH3 domains structural families.[7,36] Here and in the accompanying manuscript, we report how the knowledge of sequence conservation that we obtained by sequence perturbation is combined with studies of the kinetic and thermodynamic properties of point mutants of Raf RBD, allowing for exploration of the relationship between sequence conservation, folding and stabilization of native structure. The *de novo* design or redesign of natural proteins has shown that the absence of selective pressure for function leads to hyper-stabilized mutants and conversely, that natural proteins are slightly sub-optimal for stability.[37,38] Accordingly, the results presented below suggest that the stability and folding rate of the Raf RBD is not optimized and that this could be linked at least partly to conservation of residues for binding to *h-ras*.

## Results and Discussion

The massive sequence perturbation experiment on Raf RBD allowed us to construct an experimental sequence-positional entropy profile so that we could establish the residue conservation at each position degenerated.[15] The agreement of this experimental entropy profile with that obtained for proteins sharing the ubiquitin-like topology aligned according to their secondary structure is striking, particularly in the correspondence of local minima in entropy (Materials and Methods, Figure 1(b) and Figure S1 in Supplementary Data). The experimental entropy profile is also in good agreement with the theoretical prediction from the Conservatism of Conservatism database‡ (CoC) of the Raf RBD structure and a theoretical study based on sequence alignment and computer simulations on Raf RBD and ubiquitin.[5,39] These observations suggest that despite its intrinsic limitations the segmental sequence perturbation allows for extrapolating meaningful structural information. Furthermore, the analysis of the specific bias in occurrences of amino acids in the experimental *versus* alignments of structural analogues or functional homologues can be used to define the sequence space constraints and discriminate between their structural or functional origins (Figure 1(c)). It is clear from Figure 1(b) and (c) that the regions with the lowest entropy and the strongest bias involved mostly hydrophobic positions conserved in the β-grasp ubiquitin-like topology. The sequence conservation in the hydrophobic core suggests a bi-layer organization of the hydrophobic core, composed of an inner and outer layer. The classification of a residue as inner or outer core is based on its entropy and localization in the native structure. In addition, we identified a subgroup of residues (e.g. I58, S77, C81 and C96) that displayed a predominant selection for non-wild-type (wt) amino acids. Specific amino acid selections also revealed that important topology-defining residues were conserved, particularly obvious in some β-turns and in the α-helix. Hence, we examined sequence biases

† http://scop.mrc-lmb.cam.ac.uk/scop

‡ http://kulibin.mit.edu/coc/index.html

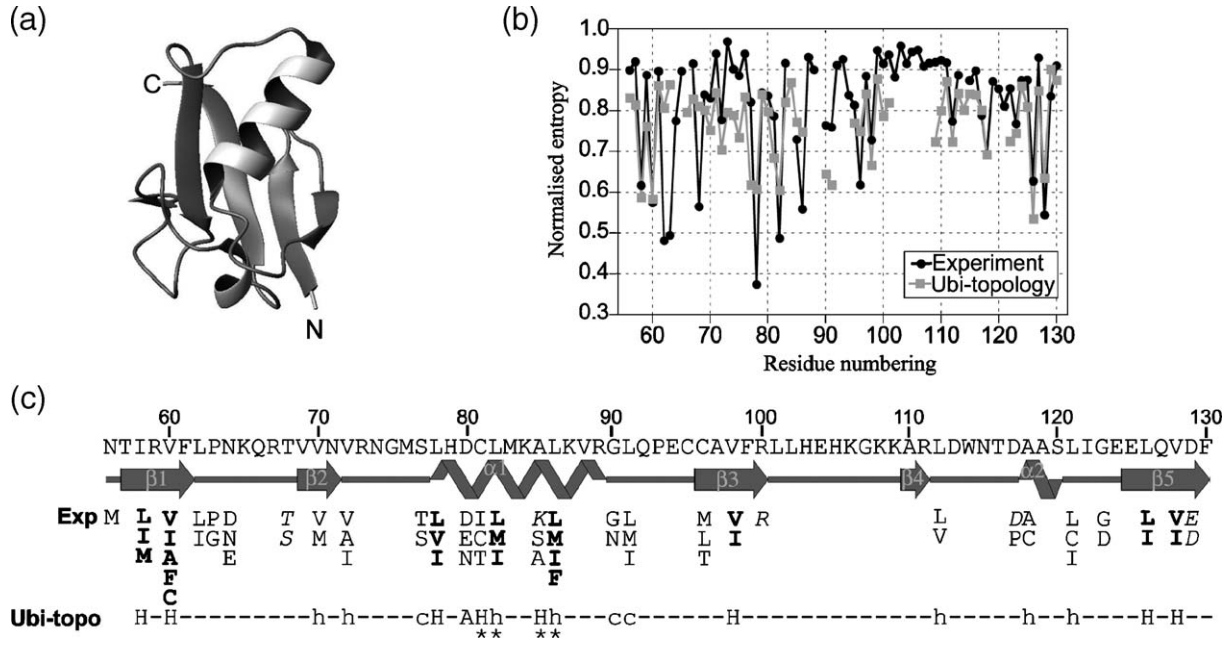


**Figure 1.** The Raf RBD model and insights obtained from sequence perturbation experiments. (a) Tertiary structure of Raf RBD. (b) Comparison of normalized entropy of the Raf RBD obtained experimentally *versus* 54 naturally occurring proteins displaying the ubiquitin-roll topology. (c) Primary and secondary structure of Raf RBD. Below the secondary structure is indicated the signature sequence of the Raf RBD, depicted as a serial group of amino acids preferentially selected at a given position during sequence perturbation experiment (Exp). The positions in bold have very low entropy both in the experiment and sequence alignments of proteins with the ubiquitin-roll topology. The positions in regular and italic lettering indicate, respectively, average entropy or conservation specific to Raf RBD. The consensus positions within the ubiquitin-roll topology (Ubi-topo) are also indicated (h, hydrophobic; c, helix capping; capital letters indicate higher conservation). Positions marked by stars in the major α-helix indicate discrepancy between the consensus Exp and Ubi-topo due to variation in its arrangement over the β-sheet.

for more subtle selection, including propensities for wt secondary structure in the sub-segments of the primary structure.

### Specific conservation of wt secondary structure propensity in segments of the β-grasp ubiquitin-like topology

An key question concerning the relative amino acid occurrences in proteins sharing the same topology is the role of local factors such as secondary structure propensity imposed in the definition of the sequence space, folding mechanism, thermodynamic stability and structure prediction. Several lines of experimental evidence argue for the presence of fluctuating elements of secondary structure in the denatured state of proteins,[40–42] including for CI2 a very well studied two-state folder.[43] These types of transitory structures could constrain the conformational search early in the folding process, thus generally agreeing with the sequential model for protein folding.[44,45] Based on this theoretical background, Rose and co-workers have proposed that secondary structure content could be used to predict the secondary structure elements able to fold in isolation and the folding rate of specific polypeptides.[46,47] Interestingly, Rosetta the most successful *de novo* design and structure prediction algorithm, uses a library of short structural segments to find local matches with the target sequence as the initial step in the tertiary structure prediction process.[48]

To explore the issue of secondary structure propensity conservation, we used the scale of Koehl and Levitt to compare the profiles of average propensity for α-helix and β-strands at all positions varied in the sequence perturbation of Raf RBD *versus* the proteins in the alignment of β-grasp ubiquitin-like topology (Figure 2(a) and Materials and Methods).[49] Overall, the patterns of propensities for α-helix and β-strand share several similarities in the experimental data *versus* natural ubiquitin-roll topology sequence alignments. In the experimental dataset for example, segments corresponding to the first β-strand (T57-F61) and the major α-helix (L78-R89) show strong preference for amino acids with high propensity for these wt secondary structures. In contrast, β2 and β5 sequences observed in the sequence perturbation studies showed low propensities for the appropriate secondary structure except at core positions. Some regions, β4 and α2 (A110-R111 and A118-S120, respectively) in particular, showed low overall propensity for wild-type secondary elements. Rose and colleagues have proposed that secondary structure elements with the highest propensity for the native structure could form early in the folding process.[46] Accordingly, the amino-terminal β-hairpins in Raf RBD and ubiquitin

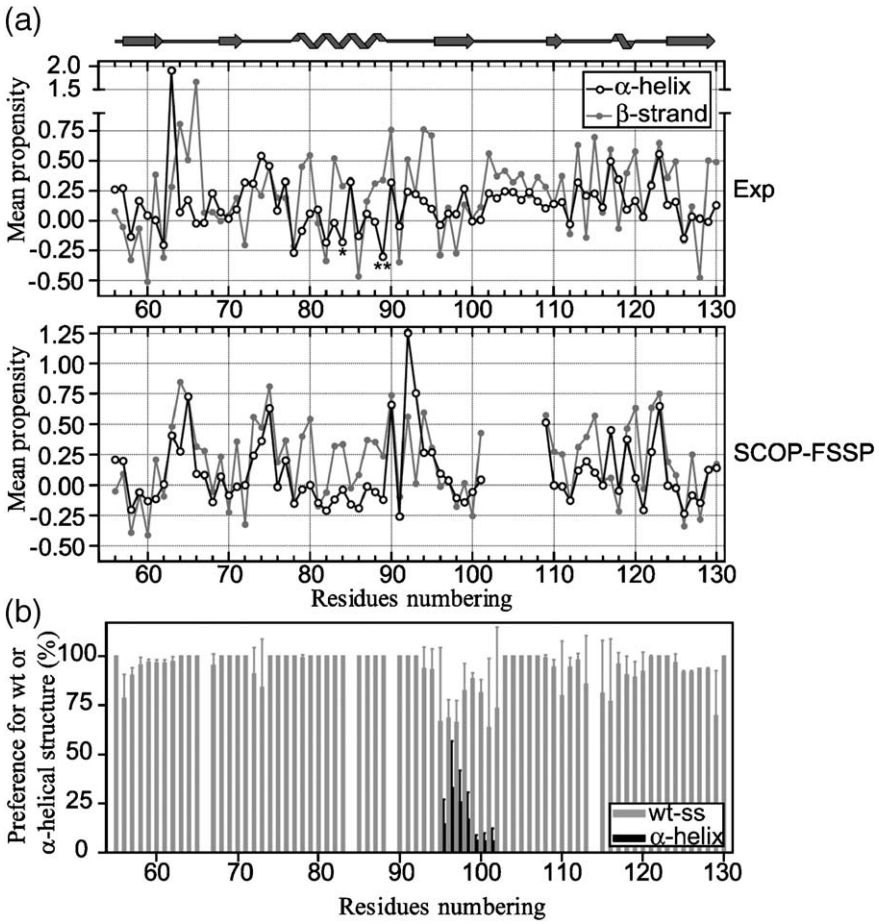


**Figure 2.** Segmental secondary structure conservation deduced from the sequence perturbation experiments: propensities and secondary structure predictions. (a) Mean propensities for α-helix and β-strand of the amino acids observed on a residue-by-residue basis in the sequence perturbation experiments and in the β-grasp ubiquitin-like alignment are shown in the top and bottom panel, respectively (Materials and Methods). Note that residues Q66, K84, R89 and W114 were not degenerated in the main experiment. Nevertheless, K84 (*) and R89 (**) showed very good conservation for amino acids with high propensity for α-helix in a separate perturbation experiment. In the case of position 89, arginine was observed in all mutants obtained.[15] (b) The mean percentage of wt secondary structure that is conserved, according to secondary structure prediction algorithms (e.g. PSIPRED, PHD and PROF), at each position varied of Raf RBD mutants obtained during the sequence perturbation experiments (■). Among the core elements of the structure, only a segment corresponding to β3 (e.g. segmental library 8[15]) showed significant reduction in wt secondary structure prediction. In this region a low but significant percentage of positions are predicted to switch to α-helical conformation (■). Further analysis of the predictions shows that 6.2(±2.8)% of clones for the region C95–L102 had at least four consecutive residues in α-helical conformation (data not shown). The gapped positions indicate the unperturbed residues mentioned in the legend for (a) (Materials and Methods).

are the most native-like region in the TS according to Φ-value analysis.[16,50] Furthermore, the ubiquitin β-hairpin was found to fold in isolation.[51] Other evidence obtained by NMR and single-molecule force spectroscopy techniques on this RBD structural analogue argue in favor of the formation of various highly structured non-native states that would be compatible with a sequential model for folding, but all are consistent with a well stabilized amino-terminal β-hairpin.[52–56]

In order to confirm that the sequence variants isolated through the sequence perturbation strategy adopt the wt secondary structure, we assessed the secondary structure of Raf RBD variants using three of the more accurate secondary structure prediction algorithms available (e.g. PhD, PROF and PSI-PRED) to calculate the average percentage of variants adopting the wt secondary structure at each position (Materials and Methods). The overview of Figure 2(b) reveals that the wt secondary structures are largely dominant with variations appearing at the margins of secondary structure elements as in β1 and β5, but more profoundly at most residues of β3 (e.g. C96-L101) for which a small fraction of the variants obtained are predicted to switch to α-helical conformation (black bars in Figure 2(b)). In fact, 6.2(±2.8)% of clones are predicted to have at least four consecutive residues in α-helical conformation between C95-L102 (data not shown). As apparent from the error bars, the algorithms are not perfectly consistent in this region, with PSI-PRED and particularly PROF predicting a higher frequency of α-helix. We also noticed that the identity of variants showing the putative secondary structure switch are far from being perfectly matched among the three algorithms (e.g. PSIPRED *versus* PROF ≈50% and PHD *versus* PROF or PSIPRED ≈0). It is noteworthy that immediately before β3, there is a tight β-turn reminiscent of a single turn α-helix; the P93 and E94 could form the N-cap residue of a putative α-helix. We hypothesize that this element of secondary structure could confound the prediction algorithms, which have a success rate of approximately 72–78% for globular proteins, or favor the induction of a true secondary structure switch. In nature, the evolution of novel protein folds from a template protein gene is dependent on the accumulation of

several single point mutations over a long time and on several mechanisms, including addition/deletion of structural elements and circular permutation, that can induce much more dramatic effects on structure.[57] In this perspective, some elements in the structure such as isolated β-turns could represent a potential nexus between known structure and the evolution of novel protein folds from the accumulation of a relatively small number of point mutations or insertion of few amino acids.

## The hydrophobic core in the ubiquitin-related superfamilies adopts two distinct patterns

The hydrophobic core of Raf RBD has a two-layer concentric organization (Figure 3(a)). The innermost layer (inner core, red) includes the residues I58, V60, L78, L82, L86, V98, L126 and V128. The outermost layer (outer core, green) includes the residues L62, Q66, T68, V70, V72, C81, A85, R89L, L91, C96, R100, L112, W114, A118 and L121, located in the immediate periphery of the inner core. These residues showed below average entropy, although generally higher than the inner core.[15] In principle, this organization can be extrapolated to ubiquitin using the alignment of natural sequences (Figure 3(b) and Figure S1 in Supplementary Data; note that Raf RBD numbering is used throughout Figures and in the text to simplify comparison and cross-references with the alignment), but there are subtle distinctions in the arrangement of the inner core of Raf RBD *versus* ubiquitin due to differences in the register of α1 that must be taken into account.

The packing of α1 over the β-sheet is found to occur mainly following two arrangements in the β-grasp ubiquitin-like topology. For example, in Raf RBD, the inner core residues of α1 are at position *i*, *i*+4 (L82) and *i*+8 (L86), while in the case of ubiquitin the second and third residues are at position *i*+3 (V26) and *i*+7 (I30), corresponding in the secondary structure alignment to C81 and A85 of Raf RBD, respectively. Careful scrutiny of the entropy profile reveals a discrepancy in the α1 hydrophobic core, which is annotated in the consensus sequence of the β-grasp ubiquitin-like topology (Figure 1(b) and (c)).[15] The impact of this on the networks of contacts established by the hydrophobic core is schematized for Raf RBD and ubiquitin (Figure 3(c) and (d)). This graph indicates that the contacts established in the hydrophobic core between residues located in the α-helix and principally at positions of the β-sheet corresponding in Raf RBD to L62, T68, V70, V98 and L126 vary across the ubiquitin-roll topology following the mode of packing of α1. For example, L62 is more intimately associated with the inner core in Raf RBD *versus* ubiquitin. The differences in arrangement are also confirmed by Φ-value analysis for Raf RBD and ubiquitin that reveals comparable involvement in TS stabilization of *i*+4 *versus* *i*+3 and *i*+8 *versus* *i*+7, respectively.[16,50]

The β-grasp ubiquitin-like topology is arranged into 12 superfamilies according to SCOP. The inner core arrangement of Raf RBD and more frequently ubiquitin occur most frequently in most superfamilies, which are thought to be evolutionarily related to the ubiquitin superfamily. The packing adopted by ubiquitin is much more frequent in this group (Table 1). Structures classified in these five superfamilies represent half (27/54 sequences) of the sequences in the alignment of natural sequences (Figure S1 in Supplementary Data). Among the six other superfamilies, four adopt structures somewhat similar to ubiquitin-related superfamilies, but many more degenerate core packing arrangements (Materials and Methods and data not shown). The evolutionary relationship between proteins displaying dissimilar or similar α1 packing is not clear and is not reflected in the SCOP classification. The comparison of contact maps for the hydrophobic core of Raf RBD and ubiquitin highlights the insights that this kind of scheme can bring to understanding the structural organization of proteins sharing similar topological structure and could be used as a method to establish evolutionary links between structural analogues and topologies.[58]

## Is the cumulative volume of the inner hydrophobic core conserved?

The observation of a common trend in the organization of the hydrophobic core contacts network spurred the analysis of the volume distribution in superfamilies of the β-grasp ubiquitin-like topology. To calculate the volume of side-chains, we used the volume of amino acids reported by F. M. Richards.[59] Previously, we noticed that the inner core is less tolerant than the outer core to volume variation. The five ubiquitin-related superfamilies showed an average cumulative volume of 594(±49) Å$^3$ in their inner core.[15] Extending the analysis to the 42 natural sequences of Figure S1 displaying the eight canonical inner core residues (e.g. residues corresponding to I58, V60, C81/L82, A85/L86, V98, L126 and V128 of Raf RBD) reveals that the distribution of the cumulative volume follows closely a normal distribution (Figure 4(a)). The inner core cumulative volume appears to be particularly constrained in the ubiquitin-related superfamilies that are intolerant to variation in volume of greater than the equivalent of three methyl groups. In other superfamilies, the volume requirements appear to be slightly more diverse and small amino acids, such as Ala or Thr occur more frequently, as indicated by the slight negative deviations observed in the distribution. The average volume of the side-chains observed at inner core positions for the ubiquitin-related superfamilies are remarkably homogenous (Figure 4(b)), with volumes corresponding roughly to the average of Val and Leu side-chains (e.g. 70–80 Å$^3$). Interestingly, the volume variation tolerated is greater when the inner core residues are considered independently than when the inner core is considered as a whole; 8.2% and 23.4% (standard deviation/mean),

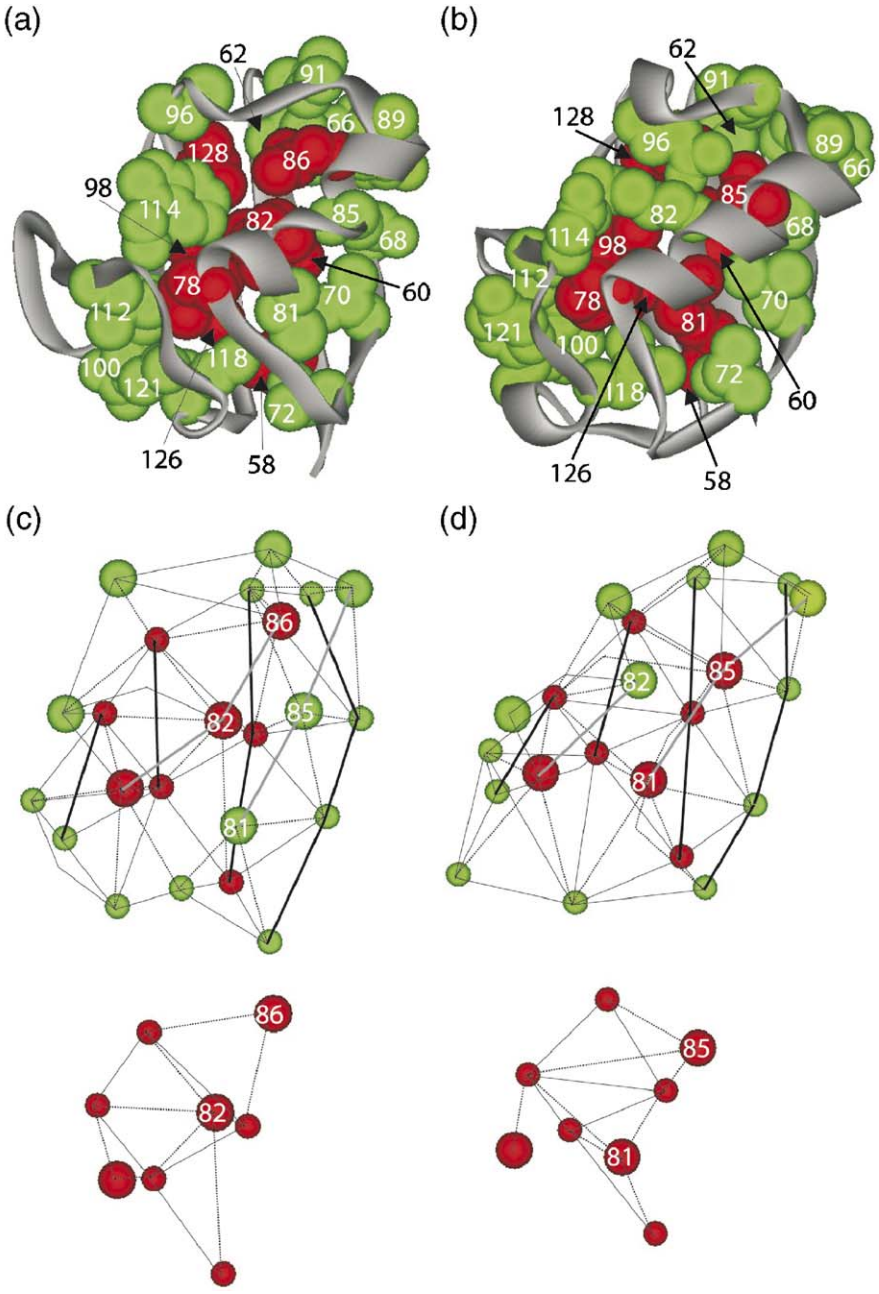


**Figure 3.** Core structural organization and side-chain contact networks: ubiquitin and Raf RBD as prototype of differing structural arrangement in the ubiquitin-roll topology. Note that the residues in both proteins are numbered according to Raf RBD sequence to facilitate comparison. (a) The hydrophobic core of the Raf RBD is organized into two layers defined as inner and outer core, readily apparent in the sequence perturbation experiment. These positions are shown on Raf RBD tertiary structure (1RFA), in red and green, respectively. (b) The homologous positions in ubiquitin are displayed on its tertiary structure (1UBI). The direct Van der Waals contacts between any two side-chains participating in the inner and outer core residues are shown by connecting $C^{\beta}$ of residues involved for (c) Raf RBD and (d) ubiquitin. These proteins consist of two surfaces, grossly defined as α1 and the β-sheet. The residues in loops are classified according to the surface in which they are integrated. The network contacts are represented in the same orientation as in the first panels. The Cβ for residues located in the α1 layer are represented by bigger spheres than the β-sheet. Thick lines connect residues that are part of the same secondary structural element (black, β-strand; grey, α-helix). In the case of α1, the residues on the same side of the helix are connected by a thick line. The thinner lines connect residues whose side-chains are in contact, whether the contact is intra-surface (continuous lines) or inter-surface (broken lines). The inner core network is shown in isolation in the bottom panel with the numbering identifying inner core residues 2 and 3 of α1. Note the variation in the arrangement of α1 over the β-sheet in Raf RBD *versus* ubiquitin.

respectively. This observation seems to suggest that there is some compensatory mechanism in the inner core that maintains the overall cumulative volume. Consis-tent with our observations, Gerstein *et al.* have observed in alignments of three protein families (globins, dihydrofolate reducatase and

**Table 1.** Amino acids observed at inner hydrophobic core residues in members of the ubiquitin-related superfamilies identified by their PDB code

| | 58[a] | 60[a] | 78[a] | 81/82[a] | 85/86[a] | 98[a] | 126[a] | 128[a] |
|---|---|---|---|---|---|---|---|---|
| 1RFA[b] | I | V | L | L | L | V | L | V |
| 1UBI | I | V | I | V | I | L | L | L |
| 1A5R | L | V | L | L | Y | F | I | V |
| 1MG8a | V | V | I | L | V | V | V | I |
| 1VCBa | L | I | V | L | V | L | V | L |
| 1M94a | V | V | V | F | L | L | L | L |
| 1J8Ca | V | V | V | F | I | L | V | L |
| 1H8Ca | L | I | L | V | V | L | L | L |
| 1JRUa | I | I | I | I | I | F | Q | L |
| 1EO6a | V | V | V | F | I | L | V | Y |
| 1EF1a | V | V | G | L | V | L | F | F |
| 1GG3a | C | V | Q | L | C | I | F | F |
| 1LFDa | I | V | A | V | A | L | F | L |
| 1E8Xa | I | I | P | I | F | L | L | L |
| 1K8Rb[b] | L | F | Y | L | L | V | L | I |
| 1L7Ya | F | I | F | V | A | I | L | L |
| 1D4Ba[b] | F | V | L | A | L | L | L | V |
| 1C9Fa[b] | V | L | L | G | F | L | L | L |
| 1F2Ri[b] | C | L | L | A | L | L | F | A |
| 1IP9a | I | F | L | L | I | L | I | V |
| 1Q1Oa[b] | F | I | L | I | I | I | I | L |
| 1FMAd | I | V | V | L | M | A | V | F |
| 1F0Za | I | F | V | L | L | L | I | L |
| 1JSBa | F | V | I | V | L | V | I | V |
| 1QF6a | I | L | P | V | I | G | L | I |
| 1JALa | Y | T | A | A | I | A | M | F |
| 1MG4 | V | F | F | L | L | I | Y | C |
| *Frequency range of amino acids (%)* | | | | | | | | |
| 40–55 | | V | | L | | L | | L |
| 20–40 | I, V | I | L, V | V | I, L | | L | |
| 10–20 | L, F | F, L | I | F, I, A | V | I, V | I, V, F | V, F, I |
| 3–10 | Y, C | T | F, P, A, G, Y, Q | G | F, A, M, C, Y | F, A, G | M, Y, Q | Y, A, C |

[a] Inner core residues as described by Campbell-Valois *et al.*[16] and Figure 4.
[b] Structures with packing *i* (res. 78), *i*+4 (res. 82) and *i*+8 (res. 86) of α-helix inner core residues. All others adopt *i*, *i*+3 (res. 81) and *i*+7 (res. 85) packing.

plastocyanin-azurin) that the cumulative volume of the hydrophobic core is better conserved than the sequence identity or the average volume observed at individual residues (respectively an average of 2.5 *versus* 13% for the three protein families) and the average volume of inner core residues was also found to be around 75 Å$^3$.[60] However they stated that the conservation of the cumulative volume was insignificant, as randomly picking the identity of the amino acid at a given core residue based on the amino acid distribution at that site in the natural sequences produced a similar result. They concluded that the inner core cumulative volume is mainly defined by the number of residues and can be explained without invoking co-variation mechanisms. It seems that randomly picking amino acids at a given position using the natural amino acid occurrence at that site indirectly embeds the co-evolutionary relationships between each position considered, particularly if sequences are highly homologous as is the case for the protein families mentioned above. Interestingly, using alignment of SH3 domains recovered from the PFAM database and structural information, we identified eight inner core residues with variation in cumulative volume of 7%, similar to what we observed in ubiquitin-related superfamilies, and a mean volume of 443 Å$^3$, which indicates that the number of residues is not sufficient to define the volume of the inner core. Nevertheless, because of the low variation in volume of hydrophobic amino acid side-chains usually observed in the core and of the variation in volume tolerated, the cumulative volume is probably a poor means of distinguishing between various protein topologies, although it could be

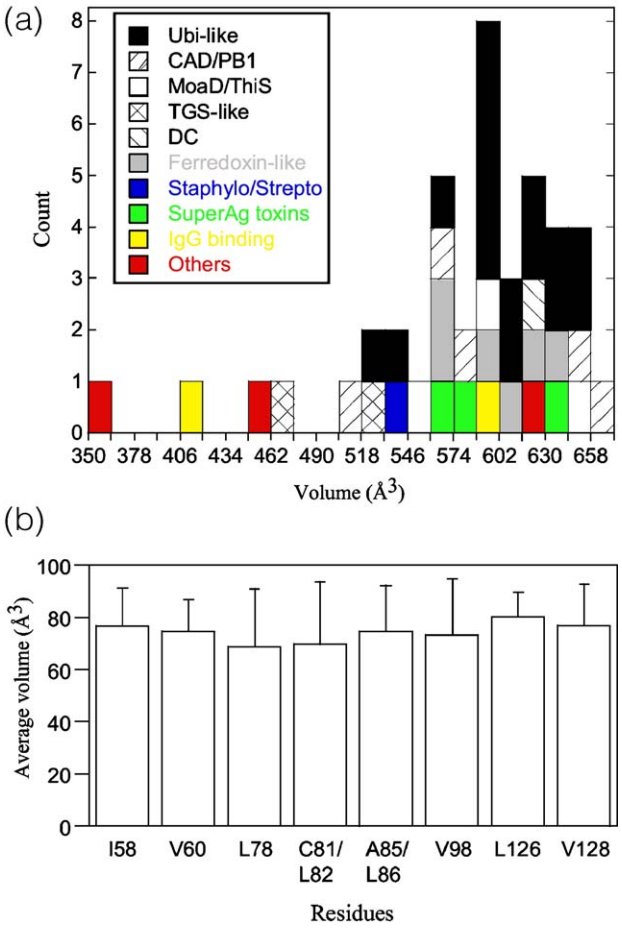


**Figure 4.** Variation of inner core residue side-chain cumulative volume in the ubiquitin superfold according to superfamily classification. The canonical inner core residues (I58, V60, L78, C81/L82, A85/L86, V98, L126 and V128) cumulative side-chain volume was calculated for each sequence included in the β-grasp ubiquitin-like topology alignment and possessing these eight residues (42 sequences from the 54 presented in Figure S1 of Supplementary Data; the remaining sequences lacked only one inner core residue, most of the time corresponding to A85/L86). (a) The cumulative volume distribution is presented in a histogram classifying sequences in their respective superfamily. The ubiquitin-related superfamilies are annotated with black and white columns. (b) The average volume and standard deviation observed at independent inner core residues for the ubiquitin-related superfamilies (27 sequences).

more significant for proteins smaller than 200 amino acid residues.[61]

We next asked whether the conservation of cumulative volume is reflected in the amino acid conservation observed at the inner core residues in the ubiquitin-related superfamilies (Table 1). As could be predicted from the low sequence identity in the alignment, the sequence requirements are flexible, with most of the positions appearing equally constrained and only four of them (e.g. 60, 81/82, 98 and 128) displaying above 40% of selection for a given amino acid. It is noteworthy, that the amino acids present in the wt Raf RBD sequence are predominant at all positions, making it a good representative model of the average inner core composition. In summary, the cumulative volume of the inner hydrophobic core appears to be conserved, despite the fact that various amino acid combinations are tolerated at every position of the inner core. In contrast, it was demonstrated that several positions in a sequence alignment grouping 266 SH3 domains are highly constrained to amino acid type.[7] This distinction could stem from the tighter evolutionary relationships in this protein family, which regroups sequences with conserved biological function. On the other hand, greater flexibility of a hydrophobic core organization could explain the high occurrence of some protein folds such as the β-grasp ubiquitin-like.[19] It is not clear that the hydrophobic core of this topology is particularly more flexible than that of other protein folds or if it is a consequence of subtle structure variation resulting from the expansive evolutionary drift in this superfold. To start delineating the biophysical meaning of the hydrophobic core organization and the contribution of other conserved positions into the formation and stabilization of the β-grasp ubiquitin-like topology structure, we performed kinetic and thermodynamic studies on Raf RBD mutants.

## Thermodynamic study on mutants of Raf RBD

On the basis of the sequence perturbation experiments and tertiary structure features of Raf RBD, 37 Ala/Gly mutations (e.g. residues are mutated to Ala, except Ala residues which are mutated to Gly) and 17 atypical mutations were introduced at selected positions. Herein, we report the thermodynamic parameters for these 51 mutants (Table 2 and Materials and Methods) and the insights that these data bring to our understanding of the native structural organization of Raf RBD. The kinetic parameters and a Φ-value analysis of the TS structural properties are reported in the accompanying article.[16]

In Figure 5(a), we show representative urea melting curves obtained for five mutants. Two different estimates of $\Delta\Delta G_{F\text{-}U}$ were calculated from the equilibrium data (Material and Methods). The $\Delta\Delta G^{0M}_{F\text{-}U}$ and $\Delta\Delta G^{Cm}_{F\text{-}U}$ are generally comparable and indicate that the quality of the data is good. However, the $\Delta\Delta G^{Cm}_{F\text{-}U}$ is more accurate, because its calculation necessitates smaller extrapolation. Correlation between thermodynamic and kinetically derived $\Delta\Delta G_{F\text{-}U}$ and $m$ suggests that the Raf RBD is a two-state folding protein and therefore that the two-state equation can be adequately applied (Table 2). Briefly, the most destabilizing mutations are concentrated in the hydrophobic core and the analysis of the thermodynamic and kinetic parameters suggests that the native state structure of Raf RBD is unaltered by mutation despite strong destabilization.[16]

## The relationship between sequence conservation and stability

The next question that we asked is whether the sequence conservation observed in the sequence perturbation experiment can be correlated either with the destabilization or reduction in folding rates induced by the Ala/Gly mutations. In order to do so, the positional entropy of Ala/Gly mutants was plotted against $\Delta\Delta G^{Cm}_{F\text{-}U}$ or $\ln k_f^{1.6M}$, respectively (Figure 5(b) and (c)). Sequence entropy correlated best with stability, rather than folding rates as indicated by regression of the linear fits ($R=0.88$ and $R=0.68$, respectively). Only P63A, T68A and C81A (●) among the 37 Ala/Gly mutants were excluded from the graph to produce these correlations. The linear correlations that included these data points as well were considerably less significant ($y=25.1-25.4x$, $R=0.75$ and $y=1.8(+4.0)x$, $R=0.56$ for $\Delta\Delta G^{Cm}_{F\text{-}U}$ and $\ln k_f^{1.6M}$, respectively). These three mutants induced less destabilization than expected from their sequence entropy. In the case of T68, it is most probably resulting from its role at the binding interface.[31] The two other residues are close to the *ras* binding surface, but are not known to be directly implicated in the interaction. It is noteworthy that the extrapolation of the linear fits shown in Figure 5 to an entropy value of 1 (e.g. a theoretical case, in which a position would display absolutely no selective pressure, meaning that all amino acids are equally well tolerated) would yield a theoretical $\Delta\Delta G^{Cm}_{F\text{-}U}$ and $k_f^{1.6M}$ of −0.7 kJ/mol and 327 $s^{-1}$, respectively. These values are within measurement errors of wt Raf RBD (e.g. 0 kJ/mol and 321 $s^{-1}$, respectively). The quality of the fits and the precision of the extrapolated values is strikingly good considering that sequences used for calculating entropy were selected using a binding assay, suggesting that function has a strong impact on conservation at only a limited set of residues. The good performance of the binding assay in this specific context may result from the wide range of $K_d$ of *ras*-RBD complex that were detectable (at least between 0.13 and 14 μM).[15] The normalization of $\Delta\Delta G^{Cm}_{F\text{-}U}$ according to volume variation, even strictly for the hydrophobic core mutations, does not improve the correlation with sequence entropy (data not shown). Hence, sequence conservation measures (e.g. positional entropy) are reliable, provided that there is sufficient sequence

**Table 2.** Thermodynamic parameters

| | $m^{a}$ (kJ mol$^{-1}$ M$^{-1}$) | $\Delta G^{ext}_{F\text{-}U}$ (kJ mol$^{-1}$) | $Cm$ (M) | $\Delta\Delta G^{0M}_{F\text{-}U}$ (kJ mol$^{-1}$) | $\Delta\Delta G^{Cm}_{F\text{-}U}$ (kJ mol$^{-1}$) | $m^{kin\ a,b}$ (kJ mol$^{-1}$ M$^{-1}$) |
|---|---|---|---|---|---|---|
| Wt | 3.8(±0.2) | −24.1(±1.3) | 6.30(±0.04) | nsap | nsap | 4.0(±0.1) |
| N56M | 3.8(±0.2) | −25.1(±1.6) | 6.59(±0.05) | −1.0(±2.0) | −1.1(±0.3) | 4.4(±0.1) |
| I58A | 4.1(±0.1) | −11.8(±0.4) | 2.83(±0.04) | 12.3(±1.3) | 13.5(±1.2) | 3.8(±0.1) |
| I58L | 4.3(±0.2) | −23.3(±1.1) | 5.47(±0.03) | 0.8(±1.7) | 3.3(±0.3) | 4.2(±0.1) |
| I58F | 4.0(±0.2) | −19.4(±0.9) | 4.87(±0.03) | 4.7(±1.5) | 5.6(±0.5) | 4.3(±0.1) |
| R59A | 3.6(±0.2) | −23.2(±1.2) | 6.46(±0.06) | 0.9(±1.8) | −0.6(±0.3) | 4.0(±0.2) |
| V60A | 3.6(±0.1) | −13.2(±0.3) | 3.70(±0.02) | 10.9(±1.3) | 10.1(±0.9) | 4.1(±0.1) |
| L62A | 3.6(±0.1) | −8.8(±0.3) | 2.43(±0.02) | 15.3(±1.3) | 15.1(±1.3) | 3.9(±0.2) |
| P63A | 3.6(±0.5) | −17.3(±2.5) | 4.73(±0.06) | 6.9(±2.8) | 6.1(±0.6) | 3.5(±0.2) |
| N64A | 3.9(±0.2) | −20.0(±1.2) | 5.15(±0.04) | 4.1(±1.8) | 4.5(±0.4) | 4.1(±0.1) |
| H2 | 4.0(±0.4) | −27.4(±2.5) | 6.77(±0.06) | −3.3(±2.8) | −1.8(±0.3) | 3.8(±0.1) |
| H2_F62L | 3.6(±0.1) | −18.5(±0.5) | 5.12(±0.02) | 5.6(±1.4) | 4.6(±0.4) | 3.9(±0.1) |
| Q66A | 3.8(±0.2) | −21.6(±1.0) | 5.71(±0.04) | 2.5(±1.7) | 2.3(±0.3) | 4.2(±0.1) |
| T68A | 3.8(±0.2) | −23.3(±1.2) | 6.14(±0.05) | 0.8(±1.8) | 0.7(±0.2) | 4.3(±0.1) |
| V69A | 3.9(±0.2) | −20.0(±0.9) | 5.07(±0.03) | 4.1(±1.6) | 4.8(±0.4) | 4.1(±0.1) |
| V70A | 4.0(±0.2) | −17.4(±0.9) | 4.32(±0.04) | 6.7(±1.6) | 7.7(±0.7) | 4.1(±0.1) |
| V72A | 4.2(±0.2) | −21.1(±1.0) | 5.08(±0.03) | 3.0(±1.7) | 4.8(±0.4) | 4.3(±0.1) |
| V72I | 3.9(±0.1) | −20.9(±0.8) | 5.32(±0.03) | 3.2(±1.5) | 3.8(±0.4) | 4.6(±0.1) |
| M76A | 3.1(±0.1) | −16.5(±0.4) | 5.26(±0.02) | 7.6(±1.4) | 4.1(±0.4) | 4.3(±0.1) |
| S77A | 4.0(±0.1) | −18.1(±0.4) | 4.57(±0.01) | 6.0(±1.3) | 6.8(±0.6) | 4.0(±0.1) |
| S77T | 4.0(±0.2) | −27.6(±1.3) | 6.97(±0.03) | −3.5(±1.8) | −2.6(±0.3) | 4.2(±0.2) |
| L78A | 4.0(±0.1) | −9.2(±0.1) | 2.28(±0.01) | 14.9(±1.3) | 15.7(±1.3) | 4.1(±0.1) |
| D80A | 3.5(±0.1) | −20.0(±0.6) | 5.76(±0.02) | 4.1(±1.4) | 2.1(±0.2) | 4.0(±0.1) |
| C81A | 3.6(±0.2) | −24.2(±1.2) | 6.70(±0.04) | −0.1(±1.7) | −1.5(±0.2) | 3.9(±0.1) |
| C81I | 5.0(±0.2) | −24.9(±1.1) | 4.98(±0.05) | −0.8(±1.7) | 5.2(±0.5) | 4.7(±0.1) |
| L82A | 4.8(±0.3) | −13.7(±0.8) | 2.84(±0.03) | 10.4(±1.5) | 13.5(±1.2) | 5.7(±0.2) |
| A85G | 4.4(±0.2) | −17.7(±0.7) | 4.00(±0.02) | 6.4(±1.5) | 9.0(±0.8) | 4.6(±0.1) |
| L86A | 4.1(±0.1) | −11.6(±0.4) | 2.85(±0.02) | 12.5(±1.4) | 13.5(±1.1) | 3.9(±0.1) |
| R89L | 4.1(±0.2) | −29.5(±1.4) | 7.27(±0.03) | −5.4(±1.9) | −3.8(±0.4) | 4.3(±0.3) |
| L91A | 3.8(±0.4) | −18.4(±2.0) | 4.59(±0.14) | 5.7(±2.4) | 6.6(±0.9) | 4.6(±0.2) |
| P93A | 4.0(±0.1) | −24.8(±0.8) | 6.17(±0.05) | −0.7(±1.5) | 0.5(±0.2) | 4.2(±0.1) |
| C95A | 4.1(±0.2) | −26.2(±1.5) | 6.38(±0.03) | −2.1(±1.9) | −0.3(±0.2) | 3.8(±0.1) |
| C96A | 3.6(±0.1) | −14.7(±0.3) | 4.05(±0.01) | 9.5(±1.3) | 8.8(±0.8) | 4.0(±0.1) |
| C96L | 3.6(±0.1) | −23.3(±0.7) | 6.40(±0.02) | 0.8(±1.5) | −0.4(±0.2) | 3.7(±0.1) |
| C96M | 4.4(±0.2) | −25.5(±1.4) | 5.73(±0.03) | −1.4(±1.9) | 2.2(±0.3) | 4.3(±0.1) |
| A97G | 4.0(±0.2) | −22.2(±1.1) | 5.49(±0.03) | 1.9(±1.7) | 3.2(±0.3) | 3.8(±0.1) |
| V98A | 3.5(±0.2) | −13.7(±0.9) | 3.97(±0.03) | 10.4(±1.6) | 9.1(±0.8) | 3.2(±0.1) |
| R100A | 3.8(±0.2) | −21.5(±0.9) | 5.62(±0.03) | 2.6(±1.6) | 2.7(±0.3) | 3.9(±0.1) |
| E104A | 3.9(±0.3) | −24.7(±1.6) | 6.31(±0.05) | −0.6(±2.1) | 0.0(±0.1) | 4.1(±0.1) |
| K109A | 3.9(±0.2) | −23.4(±1.4) | 6.00(±0.04) | 0.7(±1.9) | 1.2(±0.2) | 4.1(±0.1) |
| L112A | 3.8(±0.1) | −14.0(±0.3) | 3.67(±0.02) | 10.1(±1.3) | 10.3(±0.9) | 3.4(±0.1) |
| D117A | 3.6(±0.1) | −19.8(±0.5) | 5.51(±0.02) | 4.3(±1.4) | 3.1(±0.3) | 4.2(±0.1) |
| A118G | 3.8(±0.1) | −16.3(±0.6) | 4.27(±0.02) | 7.8(±1.4) | 7.9(±0.3) | 3.7(±0.1) |
| A118L | 3.3(±0.1) | −14.1(±0.3) | 4.23(±0.02) | 10.0(±1.3) | 8.1(±0.7) | 3.6(±0.1) |
| L121A | 4.0(±0.1) | −14.5(±0.5) | 3.66(±0.02) | 9.6(±1.4) | 10.3(±0.9) | 3.8(±0.1) |
| E124A | 4.1(±0.4) | −25.8(±2.7) | 6.27(±0.06) | −1.7(±3.0) | 0.1(±0.3) | 4.0(±0.2) |
| E125A | 3.8(±0.1) | −21.8(±0.6) | 5.73(±0.02) | 2.3(±1.4) | 2.3(±0.2) | 4.1(±0.1) |
| L126A | 4.1(±0.2) | −11.0(±0.5) | 2.71(±0.03) | 13.1(±1.3) | 14.0(±1.2) | 4.1(±0.1) |
| V128A | 4.0(±0.1) | −11.0(±0.3) | 2.74(±0.02) | 13.1(±1.3) | 13.9(±1.2) | 3.8(±0.1) |
| D129A | 3.9(±0.1) | −22.2(±0.9) | 5.38(±0.03) | 1.9(±1.6) | 3.6(±0.3) | 4.3(±0.1) |
| Δ104–6 | 3.5(±0.2) | −28.0(±1.4) | 8.03(±0.04) | −3.9(±1.9) | −6.7(±0.6) | 4.1(±0.2) |
| Δ101–8+AG | 3.6(±0.1) | −16.1(±0.7) | 4.46(±0.02) | 8.0(±1.5) | 7.2(±0.6) | 3.6(±0.1) |

See Materials and Methods for parameters description.

[a] Average $m$ and $m^{kin}$ are 3.90±0.36 and 4.07±0.41, respectively. For Ala/Gly mutants only, $m$ and $m^{kin}$ are 3.88±0.29 and 4.05±0.39, respectively.

[b] Calculated from $RT\times(-m_f+m_u)$.

diversity in an alignment, to define the importance of a given residue to the stabilization of native structure, because it is a highly context-specific parameter.

At present, there is no consensus established in the literature on the predominance of diverse biophysical parameters such as stabilization of the native structure, the folding nucleus and the folding rate on evolutionary selection. In addition, the conservation of function is likely to distort these relationships in natural proteins. However, the sensitivity of stability to mutations of Fyn SH3 that increase and decrease the volume of the hydrophobic core correlated with the frequency of the mutant amino acid in a sequence alignment of SH3 domains.[7] Recently, a theoretical study using only thermodynamic stability as a selection constraint in the simulations was sufficient to generate artificial sequences similar to natural SH3 domains at 86% of positions.[36] Interestingly, our results demonstrate experimentally that a clear relationship exists between stability and positional entropy at most positions tested. Nevertheless, a weaker

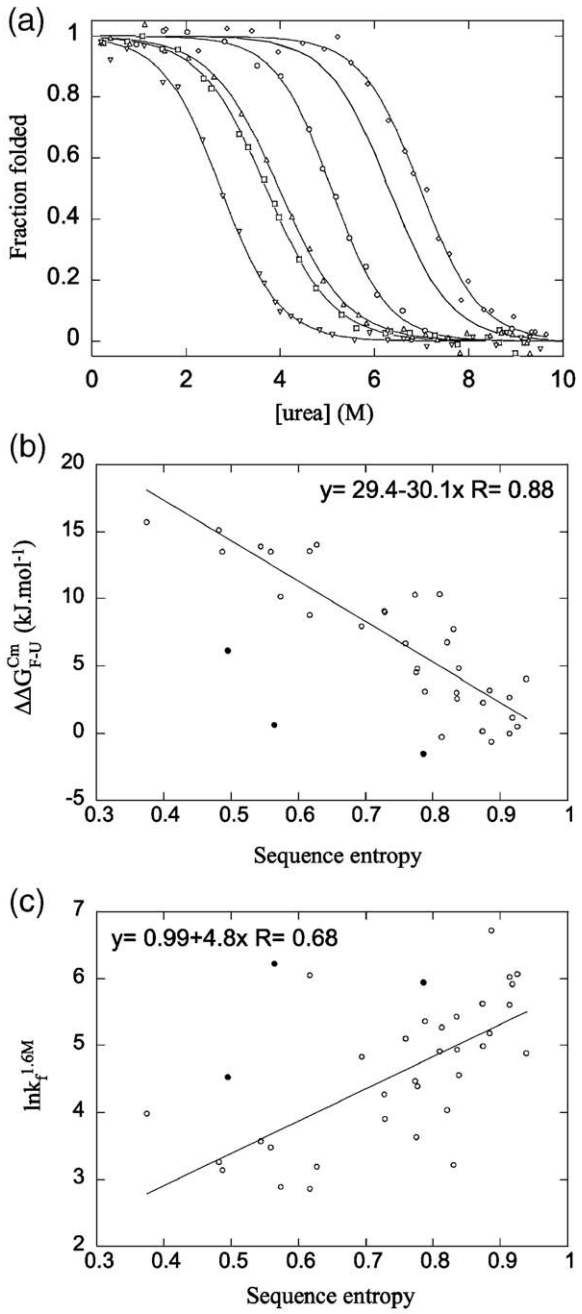


**Figure 5.** (a) Representative urea-induced melting curves of Raf RBD mutants: V60A (□), V72A (○), S77T (◊), V98A (△) and V128A (▽). Idealized wt melting curve is shown for reference (grey line). (b) Plot of positional sequence entropy *versus* $\Delta\Delta G^{Cm}_{F-U}$ induced by Ala mutation of non-Ala residue and by Gly mutation for Ala residue. The linear regression is significant ($R=0.88$). (c) Plot of positional sequence entropy *versus* $\ln k_f^{1.6\ M}$. In this case, the correlation is weaker ($R=0.68$). The mutants P63A, T68A and C81A (●) that deviated significantly were not plotted in these graphs (b) and (c). The correlations obtained with data including these three mutants were less good ($y=25.1-25.4x$, $R=0.75$ and $y=1.8+4.0x$, $R=0.56$ for $\Delta\Delta G^{Cm}_{F-U}$ and $\ln k_f^{1.6\ M}$, respectively).

correlation is also seen between entropy and folding rate. This is not very surprising given that most destabilizing mutants in this domain also induced a significant reduction in folding rate.[16] In contrast, proteins with more polarized TS should display an equivalent correlation between entropy and stability, but reduced toward folding rate as the uncoupling between the thermodynamic and kinetic parameters would be higher in this case. To verify this hypothesis and assess the potential bias introduced by the functional constraints in the conservation of residues, we sought to compare our results to those published on ubiquitin variants isolated solely on the basis of chymotrypsin resistance selection (e.g. stability).[22] Unfortunately, the low number of positions varied in that study and their concentration in the first half of the polypeptide chain precluded meaningful comparisons (data not shown). Nevertheless, using the entropy profile for ubiquitin recovered from the CoC database, we can show preferential correlation with $\Delta\Delta G_{F-U}$ and in agreement with the prediction above, poorer correlation with $\ln k_f$, most probably because ubiquitin folds through a more polarized TS than the RBD (Figure S2 in Supplementary Data).[16] Since CoC predictions are made using alignments of distant structural analogues recovered from the HSSP database, it seems that any functional bias should be minimized. By extension, the comparability of the correlations obtained with Raf RBD mutants is an additional support to previous indications that this was true in our experiments as well. As previously reported, we observed no significant correlation between positional entropy and Φ-values in Raf RBD, confirming that residues in a most native-like environment at the transition-state are not specifically conserved (Figure S3 in Supplementary Data).[33–35] The experimental demonstration of the predominant impact of stability on the selection pressure was only made possible by the high level of sequence information generated in the segmental sequence perturbation strategy. Previously, the poor sequence diversity of Raf-type RBDs and the diverging characteristics of their structure in comparison to the majority of β-grasp ubiquitin-like fold members prevented this demonstration (data not shown). Nevertheless, similarities in the entropy profiles of Raf RBD and of the β-grasp ubiquitin-like argue for comparable roles in stabilization of the native structure at the most conserved positions (Figures 1, 3 and 4). These results and observations further validate our strategy for selecting the libraries of degenerated variants and indicate that it could be applied to expand the sequence space explored by poorly populated folds.

## Map of the contribution of the hydrophobic core to the stabilization of the Raf RBD

Next, we compared the structural organization of the hydrophobic core *versus* the degree of destabilization induced by mutation of these residues to Ala/Gly (Figure 6). The most destabilizing mutations are principally located in the inner

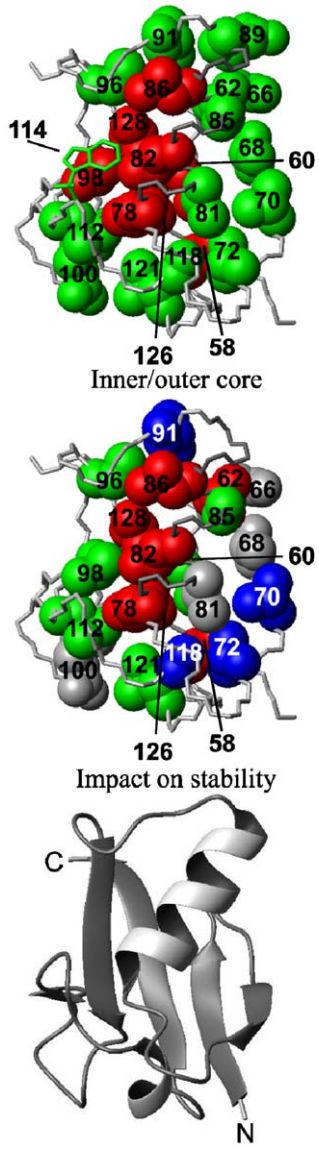


**Figure 6.** Map of the stabilizing hydrophobic core residues in Raf RBD structure. Comparison of the hydrophobic core organization determined from the sequence perturbation experiment and sequence alignments of proteins sharing the ubiquitin-roll topology (see inner core, red; outer core, green; up panel) *versus* the destabilization induced by Ala/Gly mutation of these residues (all residues mutated to Ala, but Gly mutation for Ala residues; middle panel). The residues are colored following the ratio $\Delta\Delta G^{Cm}_{F\text{-}U}/\Delta G_{F\text{-}U}$ (0–0.25, grey; 0.25–0.35, blue; 0.35–0.45, green; 0.45–0.55, orange; 0.55–0.65, red). R89 and W114 are not shown because they were not mutated for Ala/Gly. A cartoon representation of the Raf RBD is presented for reference (bottom panel).

hydrophobic core, including those located in α1, β5, plus I58 and the outer core residue L62, both located in β1. Following this set of most important residues, there is a more disparate group that includes the inner core V60 and V98, A85 in α1 and a subset of residues of the outer core (e.g. C96, L112 and L121) located mainly in the carboxy-terminal half of the domain. The β2 and remaining outer core residues dispersed on the structure play a more marginal role in the stabilization of the Raf RBD native structure. Overall, it is clear that the crucial determinant in stabilization of Raf RBD structure is located at the interface of the β-sheet and α-helix along an axis defined by the residues L78, L82 and L86, the surface of both of these topological elements forming the inner core. The similarity in entropy profile between the experimental data and the β-grasp ubiquitin-like topology (Figure 1(b)) suggests that the role of residues, particularly of the hydrophobic core, in stabilization of this topology could be conserved. In this regard, the structural distribution of stabilizing core residues in ubiquitin and its comparison to Raf RBD structural arrangement is instructive. The data on ubiquitin reveal a more prominent role for β2 and β3 and a lesser role for β5,[50] which is broadly in agreement with the variations in the hydrophobic core organization (Figure 3).

### The stability of Raf RBD is not optimal

The sequence perturbation experiment revealed a group of residues displaying high occurrence of non-wt amino acids. We sought to determine whether substitution of these residues into the wt sequence increased stability. In this study, six such variants were tested: N56M, I58L, S77T, C81I, C96L and C96M (Table 3). Only S77T showed clearly improved stability stemming mainly from lower $k_u$ (Figure 7(a) and Table 2). This mutation could possibly stabilize the structure by improving packing against the side-chain of the adjacent N115. On the other hand, N56M had higher $k_f$ and $k_u$ and displayed only marginal stabilization. The I58L, C96M and C96L mutants showed similar stability to the wt or very minor destabilization. The fact that our strategy for screening libraries of degenerated sequence of Raf RBD *in vivo* with the DHFR PCA was sensitive to mutations that disrupt the binding interface (see the text above),[15] prompted us to evaluate whether some of these mutants could have improved binding affinity. Two mutants (e.g. C81I and C96M) close to the *ras*-binding surface were selected for testing this hypothesis using an *in vitro* binding assay, but the results obtained were not confirmatory (data not shown). Alternatively, these five mutations might confer better behavior in *Escherichia coli* cells during selection. It is also foreseeable that these amino acids allowed more structural plasticity and can compensate better for destabilizing variation elsewhere in the perturbed segment.

Nonetheless, three other mutations stabilizing Raf RBD were found. The R89L that is known to disrupt the Raf RBD/*h-ras* complex (see the text above) increases stability by approximately 3.8 kJ $mol^{-1}$, mainly through improvement in folding rate. Mutant H2 recovered from the sequence perturbation experiment of the β-turn1 displayed a switch in turn type, which renders it more similar to the equivalent substructure in ubiquitin (Figure S1 in Supplementary Data and Materials and Methods).

**Table 3.** Thermodynamic and kinetic parameters for significantly stabilized mutant and double-, triple-cycle mutants

| | $m^{eq}$ (kJ mol$^{-1}$ M$^{-1}$) | $\Delta G^{eq}_{F-U}$ (kJ mol$^{-1}$) | $-m_f$ (M$^{-1}$) | $k_f$ (s$^{-1}$) | $m_u$ (M$^{-1}$) | $k_u$ (s$^{-1}$) | $\beta t^a$ |
|---|---|---|---|---|---|---|---|
| wt[b] | 3.8(±0.2) | −24.1(±1.3) | 1.22(±0.02) | 2260(±140) | 0.41(±0.04) | 0.05(±0.02) | 0.75(±0.02) |
| *Simple mutants*[b] | | | | | | | |
| H2 | 4.0(±0.4) | −27.4(±2.5) | 1.03(±0.02) | 3870(±380) | 0.49(±0.04) | 0.10(±0.04) | 0.68(±0.03) |
| S77T | 4.0(±0.2) | −27.6(±1.3) | 1.20(±0.03) | 2560(±240) | 0.52(±0.08) | 0.01(±0.01) | 0.70(±0.04) |
| R89L | 4.1(±0.2) | −29.5(±1.4) | 1.26(±0.04) | 30600(±5300) | 0.46(±0.13) | 0.04(±0.05) | 0.73(±0.06) |
| Δ104-6[c] | 3.5(±0.2) | −28.0(±1.4) | 1.32(±0.02) | 9660(±810) | 0.32(±0.10) | 0.02(±0.02) | 0.81(±0.05) |
| wt[d] | 10.9(±1.7) | −21.8(±2.7) | 3.68(±0.08) | 1100(±100) | 0.98(±0.02) | 0.31(±0.03) | 0.79(±0.02) |
| *Δ104-6: -double-triple*[e] | | | | | | | |
| Δ104-6 | 10.1(±0.8) | −27.6(±2.1) | 3.05(±0.06) | 2750(±270) | 1.01(±0.02) | 0.05(±0.01) | 0.75(±0.02) |
| Δ104-6/S77T | 9.9(±0.4) | −29.0(±1.3) | 3.25(±0.08) | 4710(±670) | 1.10(±0.03) | 0.023(±0.004) | 0.75(±0.02) |
| Δ104-6/S77T/H2 | 10.4(±0.5) | −32.9(±1.4) | 2.68(±0.09) | 7080(±1300) | 1.03(±0.03) | 0.11(±0.02) | 0.72(±0.03) |
| Δ104-6/S77T/R89L | 10.4(±0.4) | −34.3(±1.4) | 2.72(±0.07) | 13200(±2100) | 1.22(±0.03) | 0.04(±0.01) | 0.69(±0.02) |

See Materials and Methods for parameters description.
[a] Calculated from kinetic data using $\beta_t = m_f/(m_f + m_u)$.
[b] Data taken from Table 1 and the accompanying article.
[c] $\Delta G^{eq}_{F-U}$, $m_u$, $k_u$ and $\beta_t$ are not reliably measured, because the protein is too resistant to urea-induced unfolding.
[d] Data on wt Raf RBD obtained from experiments in Gdm-HCl and Tris buffer (pH 7.5) (Vallee-Belisle *et al.*[14]). Thermodynamic parameters shown here are an average of the original parameters obtained from unfolding and refolding kinetic experiments endpoint.
[e] Experiments performed in Gdm-HCl as described in Materials and Methods.

This mutant showed a modest improvement in stability, mainly through an increase in the folding rate.[16] It is usually agreed that natural proteins are not optimized for stability, because this parameter is in competition with conservation of biological function. R89L, which is the most critical residue for binding to *h-ras*,[31] is an example. *A posteriori,* the localization of this residue in the outer core, partly buried and bridging α1 to β2 suggest that it could indeed be well accommodated by a hydrophobic amino acid. In addition, α1 appears to unwind partly in a central segment between C81 and A85 due to irregularities in the H-bond patterns in the crystal structure of the complex with Rap1A charge reversal mutant *versus* the monomeric RBD structure.[30] It is noteworthy that A85 is the only α1 residue that tolerates Pro substitution in the sequence perturbation experiment.[15] Also, C81 is one of the residues, for which mutation to Ala produces lower destabilization than expected from the entropy profile (Figure 5). The unwinding of the α-helix as seen in the crystal would slightly change the packing of C81-A85 and make the R89 side-chain protrude further away from the protein interior. Therefore, we hypothesize that the suboptimal packing of the carboxy-terminal half of α1 might be involved in *ras* binding of a stable complex with *ras*. Structural comparisons of various RBD and *ras*-associated domains suggest that this phenomenon could be specific to c-Raf (e.g. for Raf RBDs). Interestingly, this RBD forms the most stable complex in simulations and experiments.[62] More speculatively, data on the H2 mutant suggests similar sequence requirements for binding to *ras* in the β-turn1 and adjacent residues that compromise the stability of the Raf RBD. In fact, this region constitutes a second major linear epitope involved in the *h-ras* binding surface and the residues mutated in H2 and R89 are adjacent in Raf RBD native structure (Figure 3).

The mutant Δ104-6 corresponds to deletion of residues E104−K106. This mutant was devised based on the observation that a-Raf and b-Raf lack these three amino acid residues. This region could confer specific functions to the c-Raf RBD as the residues deleted compose part of a putative binding site for phosphatase PP2A, similar in sequence and position in the structure to a site formally identified in casein kinase 2 α ([63] and comparison of structures adopted by residues 103−108 and 166−171 of 1RFA and 1YMI, respectively). Strikingly, this alteration produced such improved stability that the thermodynamic parameters could not be derived precisely from the urea melting curve due to the absence of a sufficiently long unfolded baseline. The Δ101-8+AG was designed based on comparison with ubiquitin, which adopts a much tighter turn in that region, but it was found to be highly destabilizing. Next, starting from the Δ104-6 background, double and triple-cycle mutants, integrating the other stabilizing mutations, were generated to determine whether they would improve thermodynamic stability of Raf RBD in an additive or synergistic manner. Equilibrium and chevron curves were generated for the various mutants, using the strong denaturant Gdm-HCl (Figure 7(b) and (c) and Table 3). First, more precise thermodynamic and kinetic parameters for Δ104-6 were obtained by this procedure; Δ104-6 is stabilized by approximately 6 kJ mol$^{-1}$ compared to the wt RBD. The mechanism by which the Δ104-6 mutant could improve stability to such an extent is not obvious. The region comprising residues L102−K108 is relatively unstructured and is among the most flexible regions of the protein according to NMR data.[64] This sequence bridges the β3 to the β4 and according to PDBsum, it forms two consecutive type IV β-turns (L102-H105 and H105-K108). Hence, the deletion of residues E104−K106 could allow for the formation of a tighter turn and less distorted β-hairpin as suggested by the structure of a-Raf (PDB

code 1WXM; no article published). This small deletion leads to an improvement in folding rate in spite of peripheral location of this segment relative to regions structured in the TS.[16] The tighter conformation of this β-turn could lead to a reduced entropic cost for loop closure and therefore would increase folding/decrease unfolding rates. Another explanation could be that the supplementary amino acids in the loop prevent this region from contributing to TS and native structure stabilization. The variation in loop size of L102-K108 suggests that the stability and kinetics for folding of c-Raf *versus* a-Raf/b-Raf might be significantly different. It would be interesting to explore how these differences could affect the normal and pathologic cellular functions that are fulfilled by the Raf genes (reviewed by Wellbrock *et al.*[29]). Predictably, the double and triple mutants (e.g. Δ104-6/S77T, Δ104-6/S77T/H2 and Δ104-6/S77T/R89L) are even more stabilized and yielded maximum improvement in folding and unfolding rates of 12-fold and 16-fold, respectively. The mutant Δ104-6/S77T/R89L displays the most improved stability, with $\Delta G_{F\text{-}U} = -34.3$ kJ mol$^{-1}$, which is −12.5 kJ mol$^{-1}$ lower than what is observed for the wt as determined in the same denaturant condition. This reduction in the free energy for folding is relatively high given that only five residues are affected. Comparable degrees of stabilization obtained by mutation of a small number of residues, which were found with experimental strategies or computer modeling specifically designed to improve stability, have been previously reported for protein-G and a cold shock protein.[65–67] The effect of mutations on Raf RBD stability seems to be additive, but not synergistic, suggesting that the mutations are optimizing different details of the native structure. Finally, the capacity of Δ104-6, double and triple mutants to bind *ras* was tested in an *in vitro* pull-down assay and competition experiment (Figure 7(d)). The Δ104-106 and Δ104-106/S77T retain a strong capacity to bind *ras*. As previously reported for wt Raf RBD, the insertion of H2 and R89L in the Δ104-106/S77T background reduces and abrogates binding to *ras*, respectively.[15,31] The specificity of the binding assay was confirmed by competition of the retention of *ras* on the resin bound Raf RBD mutants with untagged wt Raf RBD.

The ease with which we could mutate Raf RBD into a profoundly more stable form using a combination of data from the sequence perturbation experiment and literature shows that the Raf RBD sequence is not optimized for stability. Other residues involved in the Raf RBD binding surface may also be sub-optimal for stability. The degree of stabilization observed in *de novo* design experiments has indicated that the wt counterparts of designed proteins could be dramatically stabilized in the absence of selective pressure for function.[37] For

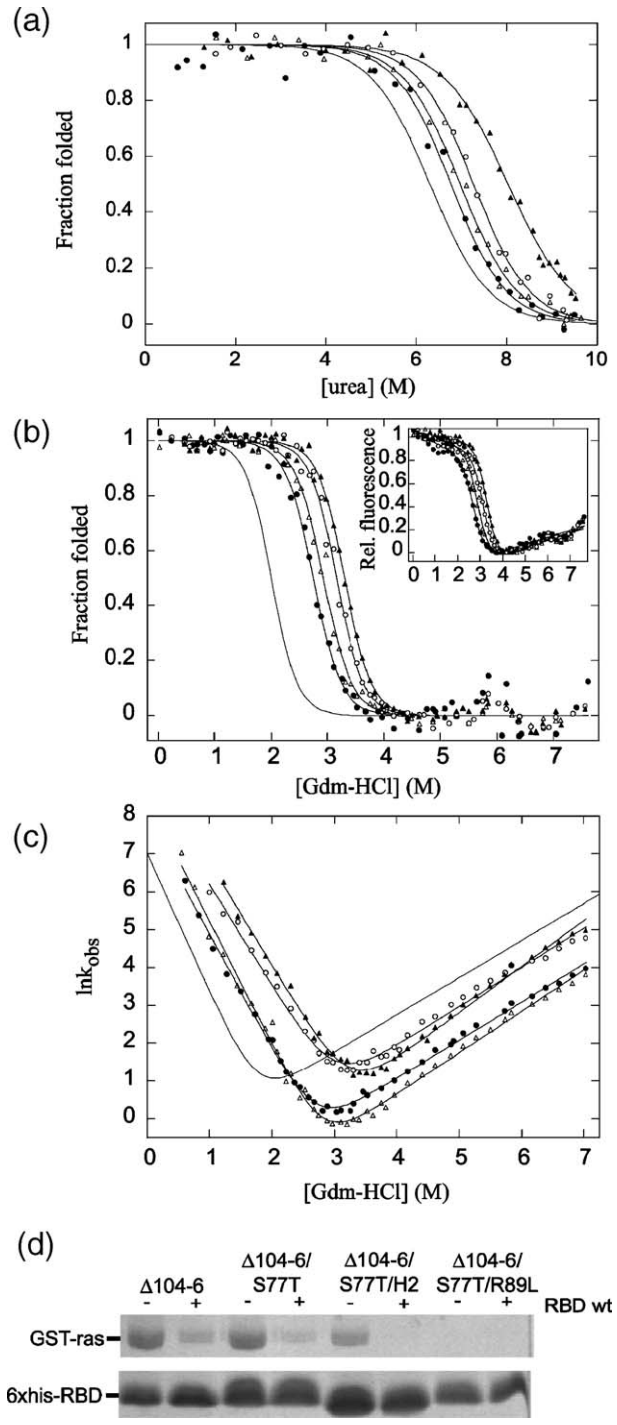


**Figure 7.** Characterization of stabilized mutants of Raf RBD. (a) Urea-induced melting curves of stabilized mutants of Raf RBD: H2 (●), S77T (△), R89L (○) and Δ104-6 (▲). (b) Gdm-HCl-induced melting curves of Δ104-6 and double and triple mutants generated from it: Δ104-6 (●), Δ104-6/S77T (△), Δ104-6/S77T/H2 (○) and Δ104-6/S77T/R89L (▲). The inset shows the raw fluorescence data for the same mutants. (c) Chevron curves for Δ104-6, double and triple mutants (symbols as in the previous panel). Modeled wt melting and chevron curves are shown in the corresponding panels to serve as reference (grey line). (d) Ni-NTA pull-down of GST-ras bound to GMP-PNP using the stabilized variants of His-tag Raf RBD and competition with untagged wt Raf RBD. The proteins were revealed by Coomassie blue staining. The picture also shows that the amount of loaded Raf RBD is similar in each lane.

example, the most drastically stabilized protein in this study, the redesigned procarboxypeptidase domain showed a 33 kJ mol$^{-1}$ diminution in $\Delta G_{F\text{-}U,}$ which corresponds to a nearly threefold improvement in stability. The absence of functional constraints that could lead to sub-optimized structural arrangements was invoked as well to explain the very high thermal stability of *de novo* designed proteins.[38]

### Conservation of a polarized distribution of charged amino acids in Raf RBD

The Raf RBD is a very basic protein with an estimated p*I* of 8.9 (e.g. 12 basic (Lys and Arg) and nine acidic (Asp and Glu) amino acids for a total length of 78 residues). A thorough mutagenesis study identified the most important residues for binding to *h-ras* on Raf RBD as R89, K84 and Q66, mostly in agreement with the insight obtained from the crystal structure of the model based on a Rap1A mutant.[30,31] As expected, because of the selection methods employed in the sequence perturbation experiments, we retrieved a clear amino acid bias at residues directly involved in binding such as Q66, T68, K84, A85 and R89.[15] The structure of the complex reveals also that the binding surface on Raf RBD includes several basic amino acids. In fact, the RBD structure displays basic and acidic patches that are segregated on opposite faces of the structure

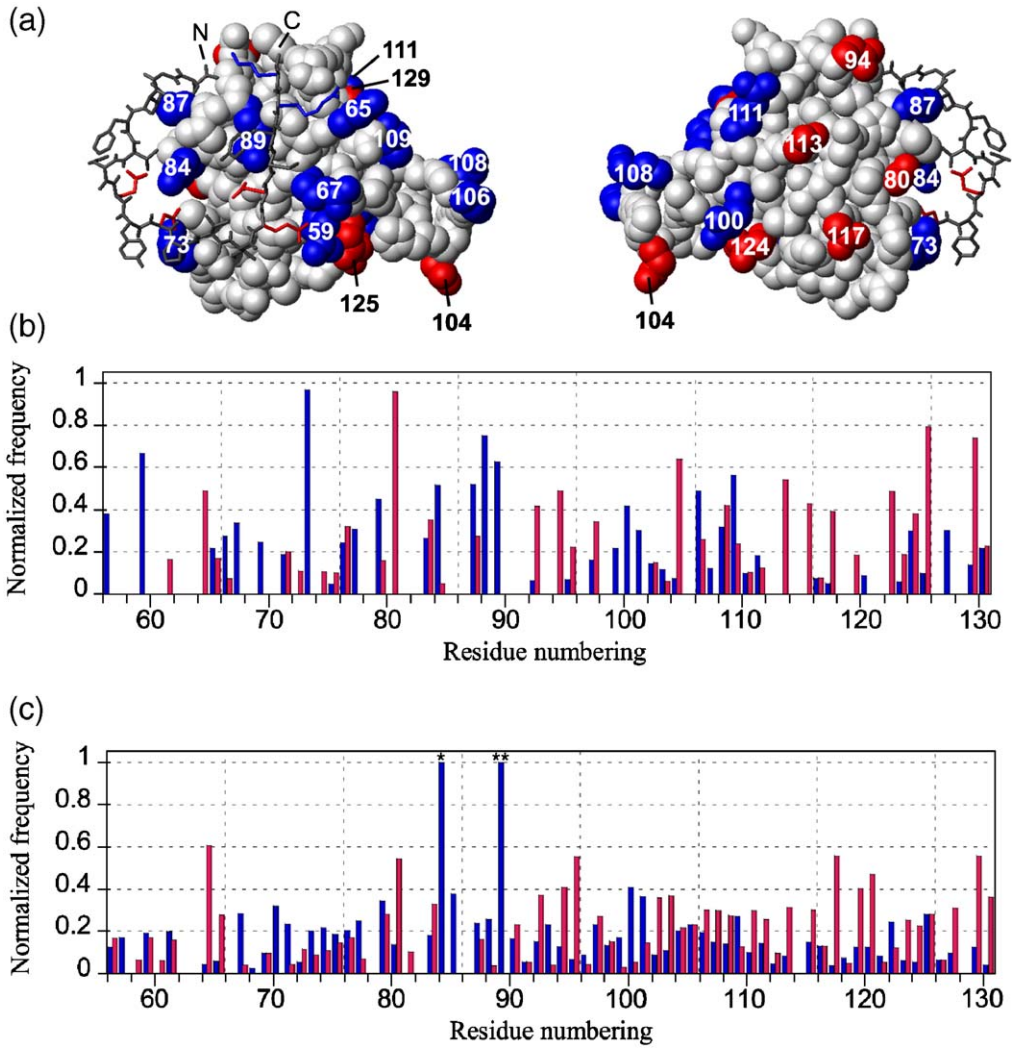


**Figure 8.** The polarized distribution of acidic and basic amino acids is conserved on the surface of the Raf RBD. (a) Basic (K, R, blue) and acidic patches (D, E, red) on c-Raf/Raf-1 RBD structure in CPK representation are shown in the left and right panel, respectively. The surfaces correspond roughly to the plane of the page and are identical image obtained by a 180° rotation around the *Y* axis. The basic patch corresponds to the GTPase binding surface as indicated by the interaction of a section of Rap1A polypeptide sequence (e.g. amino acid residues I27–K42). For the latter protein, the colored amino acid residues starting from the N termini are E31, D33, E37, D38, R41 and K42. Distribution over all positions of charged amino acids, e.g. acidic or basic, observed (b) in an alignment of Raf-type RBDs retrieved from the SMART database and (c) in the massive sequence perturbation experiment of Raf RBD. K84 (*) and R89 (**) were not varied in the main experiment, but showed average conservation for basic residues and complete conservation of arginine, respectively, when tested independently.

(Figure 8(a) left and right panel, respectively), located roughly in the N and C-terminal half of the domain, respectively. The Raf RBD displays several putative salt-bridge pairs (e.g. R100-E124, K109-D129 and R59-E125) that could be important for packing of the C-terminal half of the domain (e.g. β3 and β5) and in the establishment of long range contacts with the binding surface, although the impact on stability and folding/unfolding kinetics of mutation of putative partners was found to be inconsistent.[16]

The distribution of charged residues on the surface of Raf RBD and the fact that several of these residues showed some level of conservation in our experiments prompted us to compare the distribution of charged amino acids in an alignment of Raf-type RBD recovered from the SMART database and in the sequence perturbation experiment on c-Raf/Raf-1 (Figure 8(b) and (c)). The comparison of both histograms reveals conservation of a polarized distribution of basic and acidic amino acids. Basic residues cluster on the binding surface located in the N-terminal half of the domain, e.g. in the stretch R67–S77, K84–R89, and around R100. Interestingly, acidic residues are favored in the β-turn1 (N64–K65), the amino-terminal half of α1, in the loop constituted by the stretch G90–C95 immediately after α1 and particularly in the C-terminal half of the domain. Based on the studies discussed above, the most crucial residues (e.g. K84 (*) and R89 (**)) for binding of Raf RBD to Rap1A charge reversal mutant were not varied in the main sequence perturbation experiment in order to maximize the number of clones recovered. These residues were degenerated in independent sequence perturbation experiments, which as expected yielded a very small number of binding competent clones. This limited dataset indicated only average conservation of K84 for basic amino acids while position R89 was extremely intolerant to any amino acid switch,[15] which is in agreement with the mutagenesis data.[31] While it is expected that basic amino acids directly involved in the binding interface be conserved by our selection assay, the conservation of the acidic surface on the opposite face is more surprising. It is possible that this conservation could be due to the segment-by-segment degeneracy approach that we have utilized. However, the fact that naturally evolved Raf-type RBDs display similar charged amino acid distributions suggests that these selections are not solely the result of the experimental methodology. Also, it is noteworthy that the CAD domains of CAD and ICAD, which are classified in the β-grasp ubiquitin-like (Figure S1 in Supplementary Data; 1C9F and 1F2Ri, respectively), form a heterodimer bearing a polarized charge distribution similar to Raf RBD on the structure surface, with the basic N terminus of CAD and acidic C terminus of ICAD forming the interface.[68] The basic and acidic amino acids are also segregated in Ral GDS, but differently than in the previous examples, as can be observed from the crystallographic structure of the complex formed with *ras*.[69] These observations suggest that the polarized charge distribution has been retained by evolution in the ubiquitin-like superfamilies members involved in protein–protein interactions.

The diversity in the properties of protein–protein interfaces is very rich as observed in large sets of complexes.[70–72] Generally, the side-chains that get buried upon complex formation are as highly packed as in the hydrophobic core. The involvement of charged amino acids at buried protein interfaces reflects the less penalizing entropic contribution to protein–protein interactions than in protein folding (the role of electrostatic forces in protein–protein interactions is reviewed by Sheinerman *et al.*[73]). It has been hypothesized that the higher polarity at the interfaces of regulated dimerizing proteins reflects the necessity of the protomers to be stable and soluble on their own under physiological conditions disfavoring complex formation.[71] In fact, it is known that charged amino acids can act early in the association process (e.g. in the encounter complex) by grossly orienting and retaining together the colliding protein molecules through long range attractive and repulsive electrostatic interactions. Data on kinetics of Raf RBD association with *h-ras* are in agreement with the initial formation of such a relatively unstable complex.[74] In a key study on the importance of electrostatic interactions on the association of proteins, Schreiber and Fersht suggested that increasing favorable electrostatic interactions between protein-forming complexes would accelerate association rates by favoring a less specific TS.[75] This phenomenon could explain the moderate conservation of charges in the Raf RBD perturbation experiment at positions not crucial for *ras* binding, which would be required to ensure fast association *in vivo*. The conserved segregation of basic and acidic amino acids on the Raf RBD surface would render the basic side-chains more available for the intermolecular interaction with *ras* by avoiding intramolecular electrostatic interactions, while ensuring relatively neutral p*I* and repulsive forces in the case of non-productive encounters.

The combinations of elegant theoretical work with experiment showed that some amino acids are coupled over a long range in protein structures.[76] Specifically, it has been demonstrated that double mutant cycles of such coupled residues could synergistically affect the affinity of a PDZ domain for its substrate, despite the fact that some residues were located far from the ligand-binding pocket. These data support the hypothesis that coupled residues define a path for energy distribution across protein structure, which could play a role not only in improving binding function, but also to transmit information intramolecularly, e.g. from one face of a protein structure to another, for example the case of receptor from the cell surface to the cytoplasm.[77] Such mechanisms could be broadly implicated in signaling cascades in which they could contribute to the efficiency of proteins as biological machines and switches. The polarized charge distribution on Raf

RBD provides a tantalizing means of transmitting energy across the protein surface upon binding of *ras* at the basic binding patch. Indeed, the change in the net charge or redeployment of electrostatic interaction of Raf RBD upon binding *ras* would represent a simple mechanism to do this. However, as exemplified by the studies conducted by Ranganathan and colleagues, less obvious residue couplings could occur in the structure.

### Conclusions and perspectives

Despite several years of efforts by the scientific community, the mechanisms and interrelationships between protein sequence, function, structure stability and folding are not fully understood. Studies combining sequence alignment information and classical structural biology methods are scarce. Moreover, the use of alignments of natural sequences recovered from databases is limited to frequently occurring proteins and folds. Here and in the accompanying article, we present how the comprehensive sequence information obtained from the segmental sequence perturbation can be used to address some of the fundamental questions about protein structure and function.

The analysis of the sequence perturbation data described above has revealed that there are significant similarities in the local propensities for α-helix and β-strand between the mutated Raf RBD and an alignment of proteins sharing the ubiquitin-roll topology. Next, the determination of the thermodynamic stability and folding rate of numerous variants of Raf RBD indicates a stronger relationship of the former with sequence entropy. The Raf RBD hydrophobic core was previously described to be composed of two concentric layers, the inner and outer core.[15] The mutation of inner core residues was shown to have the most dramatic impact on thermodynamic stability and also TS stabilization, while the mutations of outer core residues had less predictable effects on stability and folding kinetics (Figure 6 and accompanying manuscript). Also, the impact of the latter class of residues is likely to vary among fold members as revealed by comparison of Raf RBD *versus* ubiquitin. However, the correlation of the entropy profiles in the inner core residues and the conservation of their structural organization and of the cumulative volume of their side-chains in the ubiquitin-related superfamilies argue for similar relationships, while the other superfamilies seem to have different properties (Figures 1, 3 and 4). In this regard, a key result is the correlation of positional entropy with stability (Figure 5), which highlights the potential utility of entropy profiles of natural sequences as a predictive tool of native structure architecture.

Finally, the combination of a few stabilizing mutations in Raf RBD indicated that its stability could be dramatically improved by affecting only five residues, leading to a 1.5-fold increase in $\Delta G_{F\text{-}U}$. The improvement in stability induced by R89L is in agreement with the hypothesis of evolutionarily non-optimized stability stemming from the necessity of maintaining binding function (e.g. more generally, any biological function). The sequence variations in the loop joining β3 and β4 between the Raf genes and the improvement in stability of the Δ104-6 are similarly interesting. These results emphasize the insights about cell signaling and function that could follow from integrating the studies of the biophysical characteristics of protein structure stabilization and of its biological functions. Also, the suboptimal thermodynamic stability in protein structures, that is to say from locally disordered regions up to fully disordered proteins, might represent a mean to expand functions accomplished and/or capacities to modulate them by providing alternative protein–protein interaction surface or accommodating post-translational modification sites (reviewed by Dyson & Wright[78]). Furthermore, novel types of allosteric sites in signaling proteins or enzymes, such as those described for the phospho-tyrosine phosphatase PTP1B and the Wiskott-Aldrich syndrome protein (N-WASP),[79,80] might be potentially generated in unstructured segments or take advantage of remodeling in these elements to perform regulatory activity.

## Materials and Methods

### Description of the β-grasp ubiquitin-like alignment

The alignment was constructed from sequences recovered in SCOP and FSSP databases. At the time this alignment was constructed there were 11 superfamilies in the β-grasp ubiquitin-like according to SCOP. The members of nine superfamilies were integrated into the alignment. Raf RBD belongs to the ubiquitin-like superfamily, which is probably evolutionarily related to four other superfamilies: CAD/PB1, MoaD/ThiS, TGS-like and Double-Cortin sequences. Recently a 12th superfamily, the TmoB-like, was added to the superfold. The sequence of the sole member of the superfamily was analyzed *a posteriori*, particularly at the level of hydrophobic core and was found to match with the observations described for the five ubiquitin related superfamilies in the SCOP database§. The four other superfamilies in the alignment are ferredoxin-like, staphylokinase/streptokinase, superantigen toxins and IgG-binding domain. Prototypes of two other superfamilies are too degenerate to be integrated in the alignment. Other proteins in the alignment were recovered from the FSSP database, but classified as "Other" topology by SCOP database. The highest sequence identity between any two sequences allowed is 35%. Any sequences that were above this threshold were not included in the alignment. Overall, the mean pair-wise identity across the alignment is 9.4% in the alignment of all superfamily sequences and 10.3% for ubiquitin-like and the four ubiquitin-related superfamilies sequences (more details concerning the alignment can be found in Campbell-Valois *et al.*[15]).

§ http://scop.mrc-lmb.cam.ac.uk/scop

### Entropy calculation

Sequence entropy was calculated using a modified version of the Shannon entropy formula as reported.[15]

### Residue-by-residue variation in mean propensity for secondary structure and acidic/basic amino acid occurrences

The normalized amino acid occurrences at each position varied in the sequence perturbation experiment, the β-grasp ubiquitin-like or natural Raf RBD alignments, as described,[15] were used to calculate average residue-by-residue secondary structure propensities and charged amino acids occurrences (Figures 2 and 8, respectively). The secondary structure propensity scale is taken from Koehl and Levitt.[49]

### Comparing secondary structure prediction of clones with wt structure

For each full-length Raf RBD variant of the 13 segmental libraries isolated during the sequence perturbation experiments,[15] secondary structure predictions were performed using these secondary structure prediction algorithms: PhD, PROF and PSI-PRED. The prediction algorithms were run with their default parameters. The wt secondary structure was also predicted using these three methods and was found to be compatible with the experimentally resolved structure. Next, for each position we calculated the mean percentage of sequence variants for which the predictive secondary structure (helix, strand or loop) was the same as predicted for the wt sequence. The standard deviation to the mean percentage was plotted in Figure 3(b). We also calculated the mean percentage of sequence variants predicted to adopt α-helix conformation in the region C95–L102 and counted the ratio of these variants that displayed at least three consecutive residues in that type of secondary structure.

### Interactions network in the hydrophobic core and inner core volume in the β-grasp ubiquitin-like

Residues of the hydrophobic core having at least one side-chain atom involved in direct Van der Waals contacts were determined by molecular graphics analysis of ubiquitin (1UBI) and Raf RBD (1RFA) structures and were linked through their $C^{\beta}$ atoms (Figure 3). The delineation of inner and outer core residues was done as described in the text. The cumulative volume of residue side-chains in the inner core was calculated using the volume estimates of Richards (Figure 4).[59]

### Mutants: description, cloning and purification

Mutants of human Raf-1/c-Raf RBD were synthesized with a variant of the ExSite™ protocol (Stratagene) using the high-fidelity Pfu polymerase. Most mutants had a single point mutation. The mutant H2 was recovered from the sequence perturbation experiment. In this mutant, residues 62–65 (Leu-Pro-Asn-Lys) of Raf RBD are replaced with the amino acids Phe-Thr-Asp-Gly. Mutant H2_F62L reverts residue 62 to the wt amino acid (Leu-Thr-Asp-Gly). Mutants Δ104-6 and Δ101-8+AG are deletion mutants. In the former case amino acid residues 104 to 106 are deleted (Glu-His-Lys). In the latter case, amino acid residues 101 to 108 are replaced by Ala-Gly, as in ubiquitin. The mutation insertions were confirmed by sequencing. The protein expressed included residues 55–132 of Raf-1 plus an amino-terminal located hexahistidine tag separated by a spacer of three amino acid residues (Ser-Met-Gly). Proteins were purified from bacterial cell lysate under denaturing conditions (e.g. using urea), on a Ni-NTA column.

### Denaturant melting curves and determination of thermodynamic parameters

The thermodynamic parameters were calculated from denaturant-induced melting curves obtained from the endpoint fluorescence (raw fluorescence at 10 s) of the unfolding traces obtained on Applied Photophysics SX.18MV stopped-flow fluorimeter (accompanying manuscript and below). All experiments were done at 25 °C, in urea and 50 mM sodium phosphate buffer (pH 7.0). The data were converted to fraction of folded protein and fit to a two-state model. Most mutants displayed minor error between the kinetic and thermodynamic estimates of $m$ (<20%), stemming principally from error in the baseline. In cases with error >20%, the thermodynamic parameters were recalculated from fluorescence melting curves performed on a Varian Eclipse spectrofluorimeter. In most cases, the discrepancies were resolved and were attributed to obvious deviations in the baselines.

The Δ104-6 was too stable to be completely denatured in urea and therefore, to derive reliable thermodynamic parameters, its denaturation and those of double and triple-cycle mutants derived from it were monitored in Gdm-HCl and 50 mM sodium phosphate buffer (pH 7.0). The thermodynamic parameters listed in Table 3 were obtained from equilibrium curves combining endpoint fluorescence from kinetic unfolding and refolding experiments (see section below for details on kinetic experiments).

$\Delta\Delta G^{Cm}_{F-U}$ was calculated using a method described previously:[81]

$$\Delta\Delta G^{Cm}_{F-U} = \langle m \rangle (Cm^{(mut)} - Cm^{(wt)}) \quad (1)$$

where $\langle m \rangle$ is the average $m$ value for all the mutants (3.90(±0.33) kj $mol^{-1}$ $M^{-1}$), $Cm^{(wt)}$ and $Cm^{(mut)}$ are the concentration of urea at which 50% of wt and mutant proteins are folded.

### Kinetics and chevron curves

The kinetics experiments were performed as described.[16] The chevron curves for stabilized mutants (Δ104-6 and double and triple cycle mutants that derived) were done under the same conditions, using Gdm-HCl as denaturant. In this case, the refolding reactions were initiated from proteins diluted in ~6.25 M Gdm-HCl. The unfolding reactions were initiated from proteins diluted in ~0.5 M Gdn-HCl.

### Ni-NTA pull-down of *ras* bound to non-hydrolyzable GTP analogs using Raf RBD

Ni-NTA pull-down of GST *ras* bound to GMP-PNP with tagged Raf RBD mutants and competition with wt untagged Raf RBD was done using the same protocol as reported.[15]

### Structure representation

The atomic coordinates of Raf RBD in its monomeric form and in complex with Rap1A charge reversal mutant, and of ubiquitin were taken from Protein Data Bank files 1RFA and 1GUA, and 1UBI, respectively. The schematic drawings were created with the MolMol and Weblab viewer (Acelrys) softwares. The Raf RBD numbering is used to identify residues in Figures, alignments and text, including for ubiquitin structural representation unless mentioned otherwise, in order to facilitate comparison and discussion of structural similarities.

## Acknowledgements

The authors thank Dr Jeffrey W. Keillor for advice and providing access to stopped-flow and fluorimeter instruments; Dr Luc Desgroseillers and members of his laboratory for granting access to electroporating device; Alexis Vallée-Belisle for discussions and data exchange. The NSERC and CIHR funded this project. F.X.C.V. is a scholar of CIHR, le programme de biologie moléculaire and the FES. S.W.M. is the Canada Research Chair in Integrative Genomics.

## Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2006.06.061

## References

1. Perutz, M. F., Kendrew, J. C. & Watson, H. C. (1965). Structure and function of haemoglobin. *J. Mol. Biol.* **13**, 669–678.
2. Rossmann, M. G. & Argos, P. (1976). Exploring structural homology of proteins. *J. Mol. Biol.* **105**, 75–95.
3. Richardson, J. S. (1977). beta-Sheet topology and the relatedness of proteins. *Nature*, **268**, 495–500.
4. Mirny, L. A., Abkevich, V. I. & Shakhnovich, E. I. (1998). How evolution makes proteins fold quickly. *Proc. Natl Acad. Sci. USA*, **95**, 4976–4981.
5. Michnick, S. W. & Shakhnovich, E. (1998). A strategy for detecting the conservation of folding-nucleus residues in protein superfamilies. *Fold. Des.* **3**, 239–251.
6. Hill, E. E., Morea, V. & Chothia, C. (2002). Sequence conservation in families whose members have little or no sequence similarity: the four-helical cytokines and cytochromes. *J. Mol. Biol.* **322**, 205–233.
7. Di Nardo, A. A., Larson, S. M. & Davidson, A. R. (2003). The relationship between conservation, thermodynamic stability, and function in the SH3 domain hydrophobic core. *J. Mol. Biol.* **333**, 641–655.
8. Kragelund, B. B., Hojrup, P., Jensen, M. S., Schjerling, C. K., Juul, E., Knudsen, J. & Poulsen, F. M. (1996). Fast and one-step folding of closely and distantly related homologous proteins of a four-helix bundle family. *J. Mol. Biol.* **256**, 187–200.
9. Perl, D., Welker, C., Schindler, T., Schroder, K., Marahiel, M. A., Jaenicke, R. & Schmid, F. X. (1998). Conservation of rapid two-state folding in mesophilic, thermophilic and hyperthermophilic cold shock proteins. *Nature Struct. Biol.* **5**, 229–235.
10. Chiti, F., Taddei, N., White, P. M., Bucciantini, M., Magherini, F., Stefani, M. & Dobson, C. M. (1999). Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nature Struct. Biol.* **6**, 1005–1009.
11. Ternstrom, T., Mayor, U., Akke, M. & Oliveberg, M. (1999). From snapshot to movie: phi analysis of protein folding transition states taken one step further. *Proc. Natl Acad. Sci. USA*, **96**, 14854–14859.
12. Guerois, R. & Serrano, L. (2000). The SH3-fold family: experimental evidence and prediction of variations in the folding pathways. *J. Mol. Biol.* **304**, 967–982.
13. Friel, C. T., Capaldi, A. P. & Radford, S. E. (2003). Structural analysis of the rate-limiting transition states in the folding of Im7 and Im9: similarities and differences in the folding of homologous proteins. *J. Mol. Biol.* **326**, 293–305.
14. Vallee-Belisle, A., Turcotte, J. F. & Michnick, S. W. (2004). raf RBD and ubiquitin proteins share similar folds, folding rates and mechanisms despite having unrelated amino acid sequences. *Biochemistry*, **43**, 8447–8458.
15. Campbell-Valois, F. X., Tarassov, K. & Michnick, S. W. (2005). Massive sequence perturbation of a small protein. *Proc. Natl Acad. Sci. USA*, **102**, 14988–14993.
16. Campbell-Valois, F. X. & Michnick, S. W. (2006). Protein engineering on Raf *ras* binding domain reveals a polarized distribution of residues with high Φ-values, but energetically diffuse transition state. *J. Mol. Biol.* (submitted for publication).
17. Larson, S. M., Di Nardo, A. A. & Davidson, A. R. (2000). Analysis of covariation in an SH3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions. *J. Mol. Biol.* **303**, 433–446.
18. Zarrine-Afsar, A., Larson, S. M. & Davidson, A. R. (2005). The family feud: do proteins with similar structures fold via the same pathway? *Curr. Opin. Struct. Biol.* **15**, 42–49.
19. Soding, J. & Lupas, A. N. (2003). More than the sum of their parts: on the evolution of proteins from peptides. *Bioessays*, **25**, 837–846.
20. Service, R. (2005). Structural biology. A dearth of new folds. *Science*, **307**, 1555.
21. Finucane, M. D., Tuna, M., Lees, J. H. & Woolfson, D. N. (1999). Core-directed protein design: I. An experimental method for selecting stable proteins from combinatorial libraries. *Biochemistry*, **38**, 11604–11612.
22. Finucane, M. D. & Woolfson, D. N. (1999). Core-directed protein design: II. Rescue of a multiply mutated and destabilized variant of ubiquitin. *Biochemistry*, **38**, 11613–11623.
23. Benitez-Cardoza, C. G., Stott, K., Hirshberg, M., Went, H. M., Woolfson, D. N. & Jackson, S. E. (2004). Exploring sequence/folding space: folding studies on multiple hydrophobic core mutants of ubiquitin. *Biochemistry*, **43**, 5195–5203.
24. Reidhaar-Olson, J. F. & Sauer, R. T. (1988). Combinatorial cassette mutagenesis as a probe of the informational content of protein sequences. *Science*, **241**, 53–57.
25. Lim, W. A. & Sauer, R. T. (1989). Alternative packing arrangements in the hydrophobic core of lambda repressor. *Nature*, **339**, 31–36.

26. Axe, D. D., Foster, N. W. & Fersht, A. R. (1996). Active barnase variants with completely random hydrophobic cores. *Proc. Natl Acad. Sci. USA*, **93**, 5590–5594.
27. Riddle, D. S., Santiago, J. V., Bray-Hall, S. T., Doshi, N., Grantcharova, V. P., Yi, Q. & Baker, D. (1997). Functional rapidly folding proteins from simplified amino acid sequences. *Nature Struct. Biol.* **4**, 805–809.
28. Kim, D. E., Gu, H. & Baker, D. (1998). The sequences of small proteins are not extensively optimized for rapid folding by natural selection. *Proc. Natl Acad. Sci. USA*, **95**, 4982–4986.
29. Wellbrock, C., Karasarides, M. & Marais, R. (2004). The RAF proteins take centre stage. *Nature Rev. Mol. Cell Biol.* **5**, 875–885.
30. Nassar, N., Horn, G., Herrmann, C., Block, C., Janknecht, R. & Wittinghofer, A. (1996). Ras/Rap effector specificity determined by charge reversal. *Nature Struct. Biol.* **3**, 723–729.
31. Block, C., Janknecht, R., Herrmann, C., Nassar, N. & Wittinghofer, A. (1996). Quantitative structure-activity analysis correlating Ras/Raf interaction in vitro to Raf activation in vivo. *Nature Struct. Biol.* **3**, 244–251.
32. Mirny, L. & Shakhnovich, E. (2001). Evolutionary conservation of the folding nucleus. *J. Mol. Biol.* **308**, 123–129.
33. Plaxco, K. W., Larson, S., Ruczinski, I., Riddle, D. S., Thayer, E. C., Buchwitz, B. *et al.* (2000). Evolutionary conservation in protein folding kinetics. *J. Mol. Biol.* **298**, 303–312.
34. Larson, S. M., Ruczinski, I., Davidson, A. R., Baker, D. & Plaxco, K. W. (2002). Residues participating in the protein folding nucleus do not exhibit preferential evolutionary conservation. *J. Mol. Biol.* **316**, 225–233.
35. Tseng, Y. Y. & Liang, J. (2004). Are residues in a protein folding nucleus evolutionarily conserved? *J Mol. Biol.* **335**, 869–880.
36. Larson, S. M. & Pande, V. S. (2003). Sequence optimization for native state stability determines the evolution and folding kinetics of a small protein. *J. Mol. Biol.* **332**, 275–286.
37. Dantas, G., Kuhlman, B., Callender, D., Wong, M. & Baker, D. (2003). A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. *J. Mol. Biol.* **332**, 449–460.
38. Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L. & Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science*, **302**, 1364–1368.
39. Mirny, L. A. & Shakhnovich, E. I. (1999). Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J. Mol. Biol.* **291**, 177–196.
40. Wong, K. B., Clarke, J., Bond, C. J., Neira, J. L., Freund, S. M., Fersht, A. R. & Daggett, V. (2000). Towards a complete description of the structural and dynamic properties of the denatured state of barnase and the role of residual structure in folding. *J. Mol. Biol.* **296**, 1257–1282.
41. Klein-Seetharaman, J., Oikawa, M., Grimshaw, S. B., Wirmer, J., Duchardt, E., Ueda, T. *et al.* (2002). Long-range interactions within a nonnative protein. *Science*, **295**, 1719–1722.
42. Religa, T. L., Markson, J. S., Mayor, U., Freund, S. M. & Fersht, A. R. (2005). Solution structure of a protein denatured state and folding intermediate. *Nature*, **437**, 1053–1056.
43. Kazmirski, S. L., Wong, K. B., Freund, S. M., Tan, Y. J., Fersht, A. R. & Daggett, V. (2001). Protein folding from a highly disordered denatured state: the folding pathway of chymotrypsin inhibitor 2 at atomic resolution. *Proc. Natl Acad. Sci. USA*, **98**, 4349–4354.
44. Ptitsyn, O. B. (1987). Protein folding: hypothesis and experiments. *J Protein Chem.* **6**, 273–293.
45. Baldwin, R. L. & Rose, G. D. (1999). Is protein folding hierarchic? II. Folding intermediates and transition states. *Trends Biochem. Sci.* **24**, 77–83.
46. Srinivasan, R. & Rose, G. D. (1999). A physical basis for protein secondary structure. *Proc. Natl Acad. Sci. USA*, **96**, 14258–14263.
47. Gong, H., Isom, D. G., Srinivasan, R. & Rose, G. D. (2003). Local secondary structure content predicts folding rates for simple, two-state proteins. *J. Mol. Biol.* **327**, 1149–1154.
48. Simons, K. T., Strauss, C. & Baker, D. (2001). Prospects for *ab initio* protein structural genomics. *J. Mol. Biol.* **306**, 1191–1199.
49. Koehl, P. & Levitt, M. (1999). Structure-based conformational preferences of amino acids. *Proc. Natl Acad. Sci. USA*, **96**, 12524–12529.
50. Went, H. M. & Jackson, S. E. (2005). Ubiquitin folds through a highly polarized transition state. *Protein Eng. Des. Sel.* **18**, 229–237.
51. Zerella, R., Evans, P. A., Ionides, J. M., Packman, L. C., Trotter, B. W., Mackay, J. P. & Williams, D. H. (1999). Autonomous folding of a peptide corresponding to the N-terminal beta-hairpin from ubiquitin. *Protein Sci.* **8**, 1320–1331.
52. Harding, M. M., Williams, D. H. & Woolfson, D. N. (1991). Characterization of a partially denatured state of a protein by two-dimensional NMR: reduction of the hydrophobic interactions in ubiquitin. *Biochemistry*, **30**, 3120–3128.
53. Briggs, M. S. & Roder, H. (1992). Early hydrogen-bonding events in the folding reaction of ubiquitin. *Proc. Natl Acad. Sci. USA*, **89**, 2017–2021.
54. Kitahara, R., Yamada, H. & Akasaka, K. (2001). Two folded conformers of ubiquitin revealed by high-pressure NMR. *Biochemistry*, **40**, 13556–13563.
55. Kitahara, R. & Akasaka, K. (2003). Close identity of a pressure-stabilized intermediate with a kinetic intermediate in protein folding. *Proc. Natl Acad. Sci. USA*, **100**, 3167–3172.
56. Carrion-Vazquez, M., Li, H., Lu, H., Marszalek, P. E., Oberhauser, A. F. & Fernandez, J. M. (2003). The mechanical stability of ubiquitin is linkage dependent. *Nature Struct. Biol.* **10**, 738–743.
57. Grishin, N. V. (2001). Fold change in evolution of protein structures. *J. Struct. Biol.* **134**, 167–185.
58. Lindorff-Larsen, K., Rogen, P., Paci, E., Vendruscolo, M. & Dobson, C. M. (2005). Protein folding and the organization of the protein topology universe. *Trends Biochem. Sci.* **30**, 13–19.
59. Richards, F. M. (1974). The interpretation of protein structures: total volume, group volume distributions and packing density. *J. Mol. Biol.* **82**, 1–14.
60. Gerstein, M., Sonnhammer, E. L. & Chothia, C. (1994). Volume changes in protein evolution. *J. Mol. Biol.* **236**, 1067–1078.
61. Liang, J. & Dill, K. A. (2001). Are proteins well-packed? *Biophys. J.* **81**, 751–766.
62. Kiel, C., Wohlgemuth, S., Rousseau, F., Schymkowitz, J., Ferkinghoff-Borg, J., Wittinghofer, F. & Serrano, L. (2005). Recognizing and defining true Ras binding domains II: *in silico* prediction based on homology modelling and energy calculations. *J. Mol. Biol.* **348**, 759–775.
63. Heriche, J. K., Lebrin, F., Rabilloud, T., Leroy, D., Chambaz, E. M. & Goldberg, Y. (1997). Regulation of

protein phosphatase 2A by direct interaction with casein kinase 2alpha. *Science*, **276**, 952–955.
64. Emerson, S. D., Madison, V. S., Palermo, R. E., Waugh, D. S., Scheffler, J. E., Tsao, K. L. *et al.* (1995). Solution structure of the Ras-binding domain of c-Raf-1 and identification of its Ras interaction surface. *Biochemistry*, **34**, 6911–6918.
65. Perl, D., Mueller, U., Heinemann, U. & Schmid, F. X. (2000). Two exposed amino acid residues confer thermostability on a cold shock protein. *Nature Struct. Biol.* **7**, 380–383.
66. Malakauskas, S. M. & Mayo, S. L. (1998). Design, structure and stability of a hyperthermophilic protein variant. *Nature Struct. Biol.* **5**, 470–475.
67. Wunderlich, M., Martin, A., Staab, C. A. & Schmid, F. X. (2005). Evolutionary protein stabilization in comparison with computational design. *J. Mol. Biol.* **351**, 1160–1168.
68. Otomo, T., Sakahira, H., Uegaki, K., Nagata, S. & Yamazaki, T. (2000). Structure of the heterodimeric complex between CAD domains of CAD and ICAD. *Nature Struct. Biol.* **7**, 658–662.
69. Huang, L., Hofer, F., Martin, G. S. & Kim, S. H. (1998). Structural basis for the interaction of Ras with RalGDS. *Nature Struct. Biol.* **5**, 422–426.
70. Lo, C. L., Chothia, C. & Janin, J. (1999). The atomic structure of protein–protein recognition sites. *J. Mol. Biol.* **285**, 2177–2198.
71. Nooren, I. M. & Thornton, J. M. (2003). Structural characterisation and functional significance of transient protein–protein interactions. *J. Mol. Biol.* **325**, 991–1018.
72. Shaul, Y. & Schreiber, G. (2005). Exploring the charge space of protein-protein association: a proteomic study. *Proteins: Struct. Funct. Genet.* **60**, 341–352.
73. Sheinerman, F. B., Norel, R. & Honig, B. (2000). Electrostatic aspects of protein-protein interactions. *Curr. Opin. Struct. Biol.* **10**, 153–159.
74. Sydor, J. R., Engelhard, M., Wittinghofer, A., Goody, R. S. & Herrmann, C. (1998). Transient kinetic studies on the interaction of Ras and the Ras-binding domain of c-Raf-1 reveal rapid equilibration of the complex. *Biochemistry*, **37**, 14292–14299.
75. Schreiber, G. & Fersht, A. R. (1996). Rapid, electrostatically assisted association of proteins. *Nature Struct. Biol.* **3**, 427–431.
76. Lockless, S. W. & Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, **286**, 295–299.
77. Suel, G. M., Lockless, S. W., Wall, M. A. & Ranganathan, R. (2003). Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Struct. Biol.* **10**, 59–69.
78. Dyson, H. J. & Wright, P. E. (2005). Intrinsically unstructured proteins and their functions. *Nature Rev. Mol. Cell Biol.* **6**, 197–208.
79. Wiesmann, C., Barr, K. J., Kung, J., Zhu, J., Erlanson, D. A., Shen, W. *et al.* (2004). Allosteric inhibition of protein tyrosine phosphatase 1B. *Nature Struct. Mol. Biol.* **11**, 730–737.
80. Peterson, J. R., Bickford, L. C., Morgan, D., Kim, A. S., Ouerfelli, O., Kirschner, M. W. & Rosen, M. K. (2004). Chemical inhibition of N-WASP by stabilization of a native autoinhibited conformation. *Nature Struct. Mol. Biol.* **11**, 747–755.
81. Jackson, S. E., Moracci, M., elMasry, N., Johnson, C. M. & Fersht, A. R. (1993). Effect of cavity-creating mutations in the hydrophobic core of chymotrypsin inhibitor 2. *Biochemistry*, **32**, 11259–11269.

***Edited by F. Schmid***

*(Received 23 January 2006; received in revised form 23 May 2006; accepted 21 June 2006)*