Massive sequence perturbation of a small protein

F.-X. Campbell-Valois*†, K. Tarassov*, and S. W. Michnick*‡

*Département de Biochimie and †Programme de Biologie Moléculaire, Université de Montréal, C.P. 6128, Succ. Centre-ville, Montréal, QC, Canada H3C 3J7

Edited by William F. DeGrado, University of Pennsylvania School of Medicine, Philadelphia, PA, and approved August 26, 2005 (received for review January 19, 2005)

Most protein topologies rarely occur in nature, thus limiting our ability to extract sequence information that could be used to predict structure, function, and evolutionary constraints on protein folds. In principle, the sequence diversity explored by a given protein topology could be expanded by introducing sequence perturbations and selecting variant proteins that fold correctly. However, our capacity to explore sequence space is intrinsically limited by the enormous number of sequences generated from the 20 amino acids and the limited number of variants likely to fold. Here we sought to test whether the sequence space for naturally existing proteins can be explored by simple, sequential degeneration of a complete set of short sequence segments of a model protein, without long-range covariation. Using the Raf *ras* binding domain as a model of a small protein capable of autonomous folding, we degenerated 72 of 76 positions of the primary structure for the 20 amino acids in segments of four to seven residues defined by secondary structure and selected the folded species for interaction with *h-ras* by using an *in vivo* survival-selection assay. The methodology presented allowed for rigorous statistical analysis and comparison of sequence diversity. The ensemble of sequence variants of Raf *ras* binding domain obtained have recaptured the diversity observed for the ubiquitin-roll topology. A signature sequence for this fold and the implication of this strategy to protein design and structure prediction are discussed.

massive mutagenesis | protein-fragment complementation assay | protein structure topology | ubiquitin superfold | Raf *ras* binding domain

The ability of polypeptides to fold into a unique native structure is remarkably robust to mutations (1–3). Thus, polypeptides sharing structural topology, particularly if unrelated functionally, can display very low sequence identity. It follows that the comparison of diverse protein sequences adopting the same structure could be used to define the sequence determinants of a specific fold, because these residues are the most likely to be conserved across multiple sequence alignments (MSA). However, this approach is limited to a minority of folds for which a sufficient number of structures having divergent sequences are available. To expand the sequence space explored by a given topology, an interesting solution is to mimic nature by introducing massive degeneracy into the amino acid sequences and select variants for their folding capacity to identify the residues that are under selective pressure. Such information could build on achievements of protein design and structure prediction algorithms (4–7).

A mutagenesis strategy aimed at exploring the potential sequence diversity of an entire protein should allow for randomization of all residues for the 20 amino acids (aa). An algorithm for performing covariation of sequence segments has been proposed, but its experimental implementation would be difficult (8). Indeed, the ability to exhaustively explore sequence space is limited by the extraordinary number of sequences generated by the combination of "*l*" randomized residues (e.g., *l* is the number of residues in the polypeptide) that increases as 20^l and the limited number of sequences that can fold into the target structure. The obvious solution is to vary fewer residues at a time as has already been done to study compensation effects, principally in the hydrophobic core, by covarying residues dispersed over the primary sequence (9–12). This method is not suitable for large-scale sequence perturbation, because of the enormous number of covariation combinations to test and technical limitations in library synthesis. Alternatively, the primary structure can be degenerated in short contiguous segments (e.g., 4–10 residues). This approach has yielded interesting insights into protein folding (13, 14), but the residue degeneracy inserted was not constant across all segments, and, thus, it is difficult to interpret the significance to folding and stability of aa selection at specific positions. In principle, such a strategy would allow researchers to tackle the sequence perturbation of a protein in a simple and exhaustive way. Strangely, no attempt has been made to entirely degenerate a protein segment-by-segment. Although it is clear that a fully exhaustive search of sequence space requires covariation of all residues, it would be possible to compare the sequence diversity obtained experimentally by segmental perturbation with those found in nature and ask whether the sequence space explored is similar. The latter exercise would require a model protein for which a large number of structures with highly diverse sequences are available.

The *ras* binding domain (RBD) of the Ser/Thr kinase Raf is composed of 78 aa and folds autonomously into a compact globular structure build by the packing of a single α-helix against a mixed β-sheet of connectivity 2-1-5-3-4 (Fig. 1*A*) (15, 16). Furthermore, the Raf RBD tertiary structure is characteristic of the ubiquitin superfold (also ubiquitin roll or β-grasp ubiquitin-like), which is one of the most common topologies in the protein universe (17). Therefore, sequences of several functional homologues (fh) and many structural analogues (sa) to Raf RBD can be retrieved from databases.

The strategy consists of creating discrete libraries of Raf RBD in which the codons of contiguous residues constituting an individual secondary structure element are randomized to allow insertion of the 20 aa and then selecting correctly folded clones by using an *in vivo* protein-fragment complementation assay (PCA) to detect protein–protein interaction (Fig. 1*A*). The experimental design and the large data set described below allowed rigorous statistical analysis of sequence diversity by building positional entropy and aa selection profiles. Finally, the experimental data are compared with MSAs of fh and sa to validate the strategy. Strikingly, these analyses revealed that the sequence diversity observed experimentally in the Raf RBD sequences approximates the sequence space explored by a MSA of sa sharing the ubiquitin-roll topology.

Methods

Generation of Experimental Libraries. Each experimental library was prepared from the corresponding deletion tagged templates by two-round PCR (*Supporting Methods* in *Supporting Text*, which is published as supporting information on the PNAS web site) by using Pfu turbo polymerase. The first round of PCRs yielded two products: one corresponds to the 5′ region of the

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: DHFR, dihydrofolate reductase; fh, functional homologues; MSA, multiple sequence alignment; PCA, protein-fragment complementation assay; RBD, ras binding domain; sa, structural analogues; SCOP, Structural Classification of Proteins.

‡To whom correspondence should be addressed. E-mail: stephen.michnick@umontreal.ca.

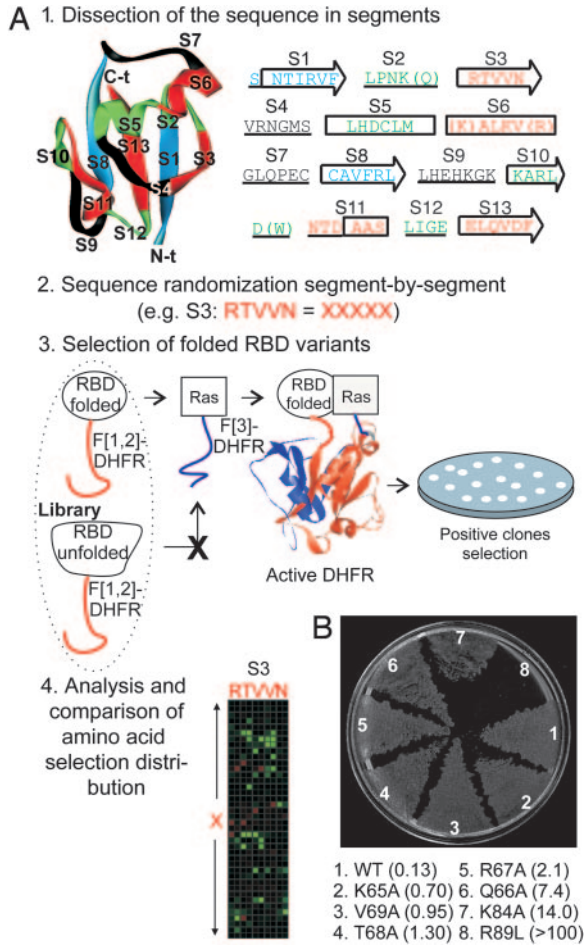


Fig. 1. Description of the sequence perturbation methodology. (*A*) Experimental strategy. (*1*) The Raf RBD is subdivided into 13 segments based on topological elements. Residues in parentheses were unvaried in the experiments. (*2*) Each segment is degenerated separately by PCR (Fig. 6). (*3*) Libraries are screened by using DHFR PCA. (*4*) Sequence diversity observed in experiments and database MSAs are compared. (*B*) Growth observed on selective media with *Escherichia coli* cells cotransformed with *h-ras* and a set of RBD mutants tethered to DHFR PCA fragments (K_d for each mutant is indicated in micromolars between parentheses below the Petri dish).

targeted segment and extending 120 bp upstream of the 5′ end of the Raf RBD cDNA and a second one flanking the targeted segment in 3′ and extending 120 bp downstream of the cDNA. Depending on the library position in the wild-type cDNA, one of the two products was generated with a loop out primer that allows reinsertion of the appropriate number of NNK codons (encoding the 20 types of aa and one terminator) into the targeted segment. All oligonucleotides were obtained from Integrated DNA Technologies (Coralville, IA) and were synthesized with hand-mixed reactants to insert the degenerated bases of the NNK codons. The PCR products generated in the first round had 18 complementary bp at their joining ends to enable the generation of the full-length degenerated cDNA through a second round PCR. The second round PCR is short (10 cycles) to maximize library representation (Fig. 6, which is published as supporting information on the PNAS web site). Detailed protocols for library recoveries can be found in *Supporting Methods* and ref. 18.

Screening Libraries with Dihydrofolate Reductase (DHFR) PCA. A fraction of pooled plasmid preparation for each library and the wt Raf RBD construct were precipitated with ethanol and dissolved in deionized water. Then, DNA concentration is estimated from OD_{260} and the volume adjusted to obtain similar concentration (e.g., 100 ng/μl). Routinely, 150 ng of the precipitated plasmid was electroporated into 60 μl of BL21 pREP4 strain carrying a vector that allows for expression of *h-ras*-DHFR [3] fusion under control of lacIq repressor. After incubation under vigorous shaking of electroporated cells in 2 ml of SOC medium during 30 min at 37°C, they were washed and plated on selective medium (as in ref. 19, except that thymine was replaced by 30 μM thiamine and 800 μg/ml casamino acids, and trimethoprim concentration was increased to 10 μg/ml). Petri dishes were incubated during 36 h at 30°C. For statistics, a dilution (wt plasmid: 1×10^{-4} and libraries: 1×10^{-3} are used) of the transformation reaction was plated on a separate Petri dish, and resulting colonies were counted (Table 1). The plasmids of selected clones were prepared in library pools by harvesting all colonies from Petri surface (18). Selected clones were sequenced, and nonredundant sequences were aligned with the wt Raf RBD (Table 2, which is published as supporting information on the PNAS web site).

Table 1. Statistics concerning the synthesis and screening of the 13 degenerate libraries

Libraries	Theoretical size ($\times 10^6$)	Real size ($\times 10^6$)	% of positives	No. of clones sequenced
S1	85.8	5.9	0.18	65
S2	0.19	2.2	0.28	61
S3	4.1	2.2	0.55	70
S4	85.8	2.4	0.25	118
S5	85.8	2.5	0.07	72
S6	0.19	0.9	0.7	81
S7	85.8	1.2	0.04	67
S8	85.8	2.3	0.3	91
S9	1800.1	3.1	1.29	74
S10	4.1	1.6	0.54	82
S11	85.8	3.2	0.39	72
S12	0.19	1.7	0.37	69
S13	85.8	1.9	0.16	64

Positional Entropy. Shannon entropy is calculated by using Eq. **1** (19):

$$S = -\sum p_i \ln p_i / \ln L. \qquad [1]$$

For experimental entropy calculation, L was fixed to 20, because we considered every switch of aa at a given position as a mutation. Before calculation of the entropy, the frequency of each aa (p_i) was corrected according to the bias introduced by the NNK codon (*Supporting Methods*). A pseudo aa was added for the entropy calculations of natural sequence alignments to account for gap occurrences. Positions displaying >25% gaps in the natural sequences MSAs were not displayed in the graphs to avoid distorting the entropy profiles. The relative entropy scores range between 0 and 1, corresponding, respectively, to total conservation and maximal exploration of sequence space (20 aa occur at same rate).

We hypothesized that N experimental sequences equal to that sampled in the screen would represent sufficient sequence space coverage to assure that most tolerated substitutions at a residue have occurred. To verify this assumption, a simulation was performed by using the following algorithm programmed in C++: n sequences were randomly selected from a complete set of N sequences, where $n = 20, \ldots, N$. For each n sequences, entropy was calculated according to Eq. **1**. Construction of a

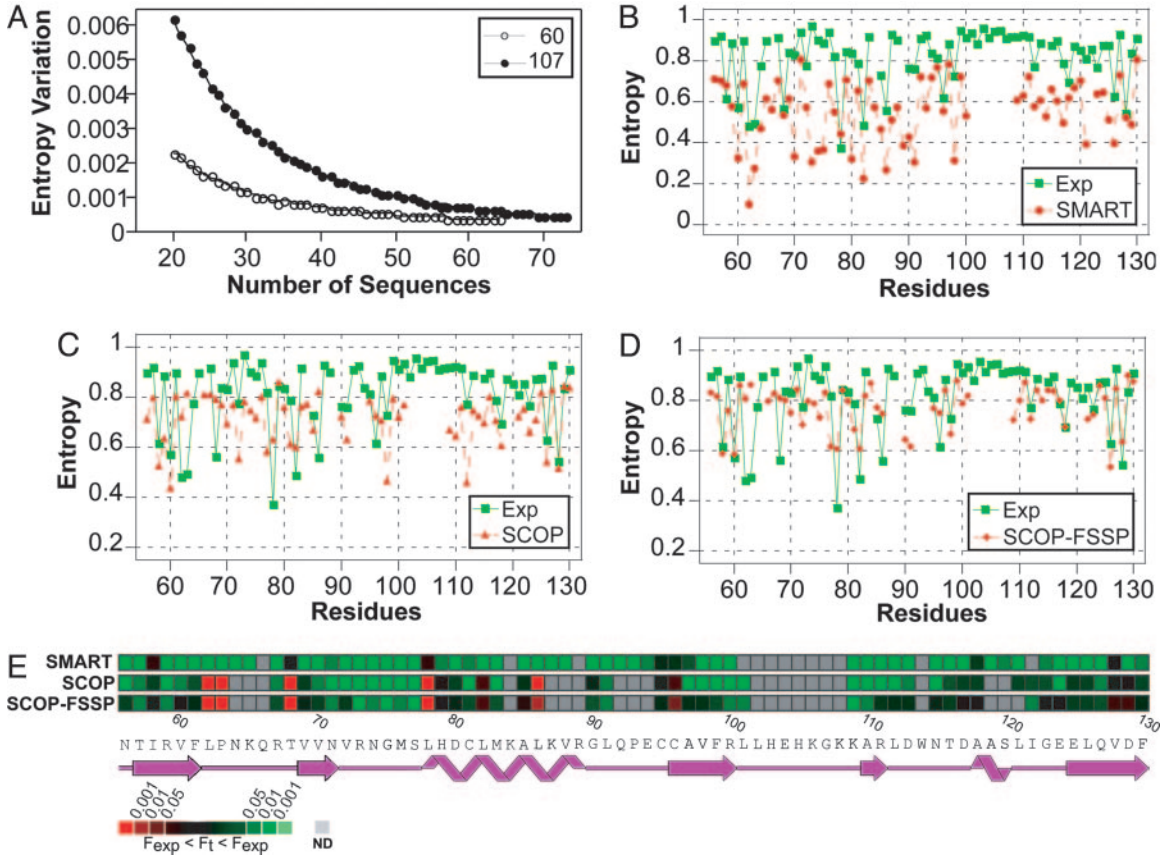


Fig. 2. Validation of experimental data set and comparison of entropy profiles. (*A*) The variation in mean entropy calculated for two successive values of *n* (*n* and *n* + 1) is plotted for V60 and G107 representing, respectively, ''low'' and ''high'' entropy position. Next, the entropy profile obtained experimentally is plotted against the entropy profiles calculated for three MSA of natural sequences: Raf RBD fh (SMART) (*B*), close sa (SCOP) (*C*), and close and distant sa [SCOP-Families of Structurally Similar Proteins (FSSP)] (*D*). (*E*) The experimental entropy profiles and the three database MSAs are compared by z score after the color scale.

subset of *n* sequences and entropy calculation was repeated 1×10^6 times, and the average entropy was calculated.

Comparison of Entropy Profiles and aa Selection by Standard Error of Proportion. The entropy profile and aa selection comparisons were calculated according to the standard error of proportion formula:

$$Z = \frac{Fe_{posX} - Ft_{posX}}{\sqrt{Ft_{posX} - (1 - Ft_{posX})/N}}. \qquad [2]$$

In which N is the number of sequences in the sample, Fe_{posX} the frequency of an aa observed experimentally at a given residue and Ft_{posX} its theoretical frequency. A positive z score means that the entropy is higher in the experimental data versus either of the natural sequence MSAs or that an aa experimental occurrence is higher than expected by chance (Figs. 2*E* and 3). The opposite is true for negative values (*Supporting Methods*).

Results

Synthesis and Screening of Libraries. The libraries were synthesized by PCR. To avoid unwanted bias for wt codons that can be introduced by this technique, the sequence to be targeted for degeneracy was removed from the wt Raf RBD cDNA before an amplification reaction that allows for insertion of the appropriate number of NNK codons (*Methods*).

The simplest way to evaluate the capacity of a sequence to fold into a target structure is to screen for the folded protein species ability to bind a known protein partner or ligand. We previously reported a simple survival-selection assay to screen libraries for protein–protein interactions based on the DHFR PCA in *E. coli*, which can detect the interaction between *h-ras* and the c-Raf (thereof Raf) RBD (20, 21). The residues of the Raf RBD directly involved in the formation of the interaction with *h-ras* were identified meticulously in a mutagenesis study (22). Several of these mutants were used to assess the sensitivity of the DHFR PCA assay to detect formation of the complex with *h-ras*. Colony formation was observed for variants of the Raf RBD displaying dissociation constant (K_{d}) between 130 nM and 14 μM. However, the R89L mutation, known to disrupt binding to *h-ras* (22), does not allow growth in this assay, and this residue shows low tolerance to mutation (Fig. 1*B*; see also Fig. 7 and Table 3, which are published as supporting information on the PNAS web site). Thus, we concluded that the DHFR PCA is sensitive enough to detect clones that fold and bind to *h-ras* with biologically relevant affinities. Based on the experimental strategy described above, we synthesized and screened 13 independent libraries (Fig. 1*A* and Tables 1 and 2).

Validation of Experimental Data Set Size. We first determined whether the interpretation of the experimental data could be biased because of the limited sampling of sequences in this study (Table 1). To test this assumption, we devised an algorithm to evaluate how Shannon entropy changes as the number of randomly sampled sequences included in the calculation is increased (*Methods*). If the sequence coverage is reasonable, the rate of change in entropy should approach zero as sequences are added. Results suggest that for the number of sequences sampled in this

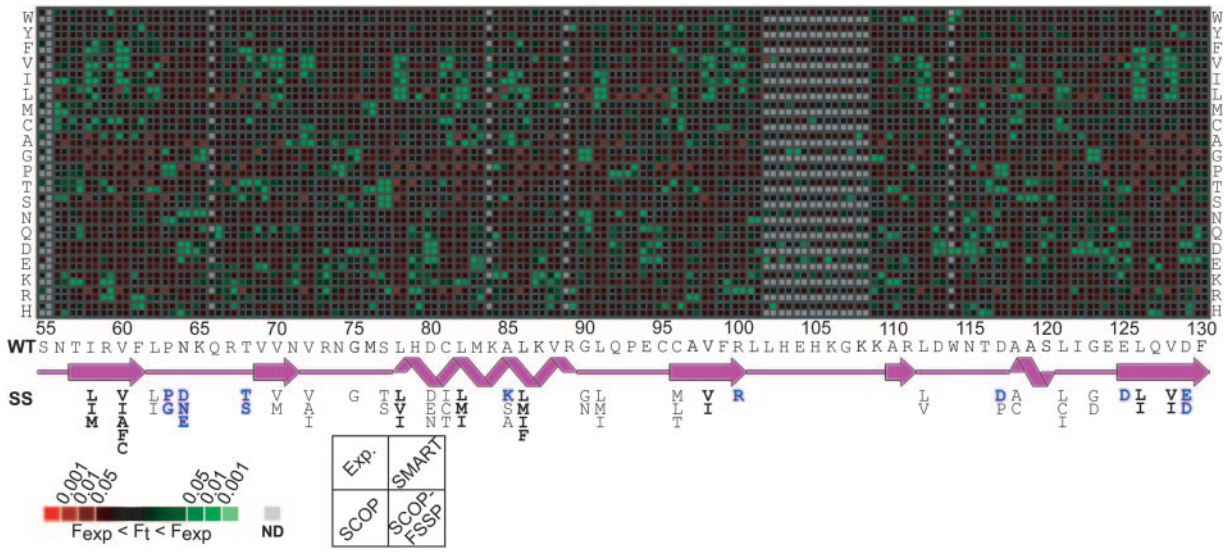


Fig. 3. Amino acid selections. The z scores of aa selections for experimental libraries and the three natural sequence MSAs are represented in a color-coded matrix after the color scale shown. Under each residue, displayed left to right on top of the matrix, there is a column of 20 cells corresponding to the aa types. Each cell is divided into four squares (see scheme below the matrix) to facilitate comparison of the MSAs. The Raf RBD signature sequence is presented below the matrix by indicating significant amino acid selection, which are classified either as conserved broadly in the ubiquitin superfold (bold) and the ubiquitin-related superfamilies (black) or specifically in the Raf RBDs (blue) (Table 9).

study, entropy variation converged toward zero whether a residue had low (V60) or high overall entropy (G107) (Fig. 2*A*).

Comparisons of Sequence Diversity in the Libraries Versus Natural Sequences. The key validity test of our experimental strategy is to show that individual positions have sequence diversity equivalent to those found in nature, despite the fact that the rest of the polypeptide sequence is held constant during the selection process. If this premise is true, we reasoned that the positional entropy profiles of the experimental data set should reflect what is observed in natural sequences. We retrieved sequences for fh and sa of Raf RBD from databases and generated three MSA. The SMART MSA includes strict fh, whereas the Structural Classification of Proteins (SCOP) and SCOP-Families of Structurally Similar Proteins MSA include sa (*Supporting Methods* and Table 2; see also Tables 4–6, which are published as supporting information on the PNAS web site).

The entropy profiles of the experimental data set and SMART MSA reveal little similarity, except in convergence of local minima of entropy (such as V60, L62, and P63) (Fig. 2*B*), because of the very high local sequence similarity of this database MSA. Overall, the higher entropy of the experimental data set reveals that the sequence diversity generated is well above what is observed in SMART MSA. The sa MSA entropy profiles are more similar to the experimental data set entropy profile (Fig. 2 *C* and *D*). Specifically, 11 residues have entropy scores below at least one standard deviation from the mean in the experiments (Table 7, which is published as supporting information on the PNAS web site): I58, V60, L62, P63, T68, L78, L82, L86, C96, L126, and V128. On the same basis, six positions (58, 60, 78, 82, 126, and 128) correspond also to local minima in the entropy profiles of the sa MSAs. The comparisons of the experiments versus the SMART and both sa MSA entropy profiles by z score analyses reveal more positions with significant differences for the former comparison (Fig. 2*E* and *Methods*). In retrospective, these results suggest that the strategy used has succeeded in reproducing the sequence diversity observed in known natural structures sharing the ubiquitin-roll topology.

The experimental entropy profiles show that the main α-helix (spanning L78–R89) core residues (L78, L82, and L86) support less degeneracy than the core positions located in the β-sheet (Fig. 2 *B–E*). This result might arise from the importance of maintaining the α-helix core packing for binding to *h-ras* as R89, a critical residue for binding (Fig. 1*B*), is located in this region. It is also possible that the core residues in the helix are more important for folding or stability of the Raf RBD than those in the sheet. On the other hand, we cannot exclude the possibility that strong positional selection in the helix-coding sequence might be a consequence of the fact that it was varied in two discrete segments, thus restraining putative local compensations specifically crucial for helices. Nevertheless, these results do not change our general conclusions and subtle local sequence constraints could be tested by using a library in which the entire helix or various segments of it are varied (*Supporting Results* in *Supporting Text*; see also Table 6, which is published as supporting information on the PNAS web site).

Hierarchy in the Hydrophobic Core. To analyze more closely the positions under selective pressure in the experiments, the aa occurrences observed experimentally and in the natural sequence MSAs were examined residue by residue by using standard error of proportion (z score) to reveal all significant aa selection (*Methods* and Fig. 3; see also Tables 8 and 9, which are published as supporting information on the PNAS web site). The experimental data set reveals that 30 of 76 positions have a z score for at least one type of aa with a P value <0.01 (*Supporting Results* and Table 9). Of these residues, 26 show strong selection for wt aa. Further, obvious convergence between the experimental, SCOP, and SCOP-Families of Structurally Similar Proteins MSA revealed 18 of 30 residues, which reside in two structural regions. They consist of an inner core (I58, V60, L78, L82, L86, V98, L126, and V128) readily evident in the entropy profiles (Fig. 2 *B–D*) and an outer core (L62, T68, V70, V72, C81, A85, L91, C96, L112, A118, and L121), which surrounds the inner core and form contacts at the interface between the α-helix and β-sheet (Fig. 4). This hierarchy in the core is not apparent in thermal b-factor or solvent accessibility data (Fig. 8 and Table 10, which are published as supporting information on the PNAS web site). For example, residues of the outer core such as V72, C81, A85, and A118 have solvent accessibility comparable to inner core positions. Based on these observations, a signature

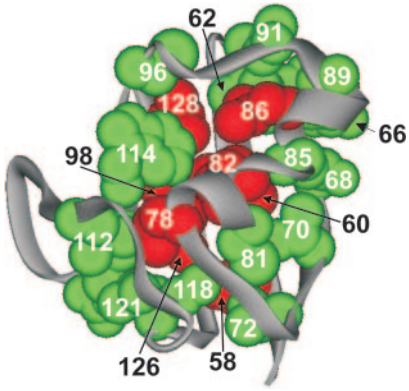


Fig. 4. Hierarchy in the hydrophobic core. Structural disposition of residues that constitute the inner (red) and outer core (green) of Raf RBD. Note that three residues for which insufficient experimental data are available likely belong to the outer core as we define it (Q66, R89, and W114).

sequence in the form of a series of significant aa selections at key residues of Raf RBD is presented (Fig. 3). The convergence of the aa occurrences observed experimentally to either of the natural sequence MSAs is used to predict the type of constraint, either functional or structural, imposed by selection at each residue of the signature sequence.

Key Topological Constraints. To further validate our data, we searched for predictable aa selections that follow from accepted principles of protein structure. For example, aa such as Pro and Gly display negative z score value in helical and β-strand segments, whereas they occur frequently in clones selected from libraries corresponding to loop or β-turn elements (libraries S2, S4, S7, S9, and S12). Also, hydrophilic residues are largely absent from hydrophobic core positions. We also observed strong selection for residues constrained by topology of the ubiquitin superfold. For example, residues S77, D80, G90, and L91 are located at the extremities of the major α-helix and show aa selection typical of helix capping motifs. On the other hand, residues P63 and N64, which form the first β-turn, represent examples of Raf RBD specific constraints. Indeed, aa selections at these positions are not as strong in the sa MSAs, because of variations in conformation and length of the matching structural segment (Figs. 2 and 3, *Supporting Methods*, and Table 4).

A Conserved Raf RBD Binding Patch for *h-ras*. Other residues show strong comparable aa selection in the experimental and fh but not in the sa MSAs (Figs. 2*E* and 3). This observation could suggest that these residues play a role in Raf RBD binding to *h-ras*. For example, observed aa selection and spatial proximity of the side chains of residues Q66, T68, R89, and A85 in the Raf RBD structure support their role in forming a critical binding surface for *h-ras* (Fig. 3 and Table 3). As discussed previously, R89 is the single residue of the Raf RBD, which is critical for binding to *h-ras*. Interestingly, Q66A and T68A decrease the affinity for *h-ras* (22), whereas A85K increase the affinity putatively by allowing for binding to a wider range of GTPase conformers (23, 24). Consistent with the latter observation, our data indicates a strong selection for lysine at position 85.

Another subset of residues, including R100, D117, E125, and D129, shows specific convergence of aa selections with the fh MSA. These residues are not known to be involved in binding to *h-ras* and are far from the binding interface. Their mutation to Ala have no or only marginal effect on the affinity for *h-ras* in an *in vitro* binding assay (Table 11, which is published as supporting information on the PNAS web site). They could possibly be involved in structure stabilization specific to Raf RBDs (SMART MSA).

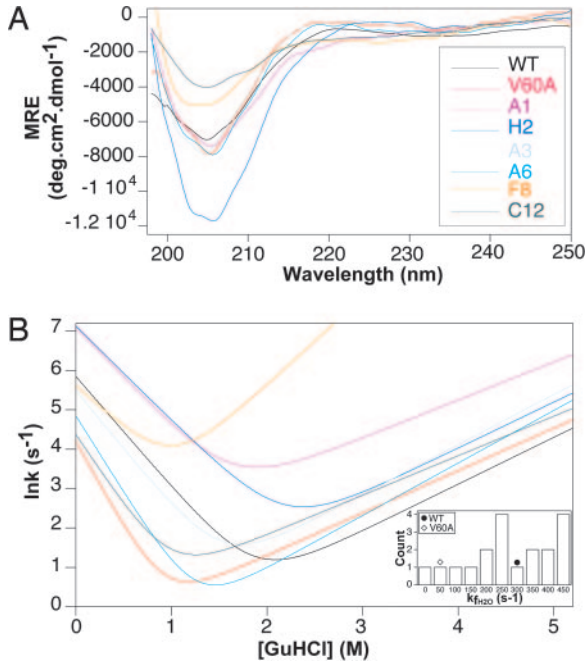


Fig. 5. Clones selected display characteristics of folded proteins. (*A*) Circular dichroism spectra for five clones from five libraries interspersed in the Raf RBD sequence. (*B*) Chevron curves obtained from folding and unfolding experiments in GuHCl for the clones presented in *A*. *Inset* shows the distribution of folding rates in water extrapolated from the chevron curves for all clones with two-state folding behavior that were studied. Note that clones with folding rate (k_f) >450 s^{-1} are grouped in a single class.

Clones Display wt-Like Folding and *h-ras* Binding. The folding and *h-ras* binding properties of Raf RBD clones were compared with the wt Raf RBD. To do so, we attempted the purification of 96 variants selected from the 13 libraries; 64 clones were purified with reasonable yield (Table 2). Far UV circular dichroism and proton NMR data suggest that the variants have conserved Raf wt structural characteristics (Fig. 5*A*; see also Fig. 9, which is published as supporting information on the PNAS web site). Moreover, similarity in the folding kinetic parameters between the variants and wt Raf RBD suggest conservation of folding mechanism despite large variation in folding rate (k_f) (Fig. 5*B*; see also Table 12, which is published as supporting information on the PNAS web site) (16). Finally, the observation that the Raf RBD variants/*h-ras* complex can be competed by wt RBD in a pull-down experiment, along with the K_d determined for some of these complexes, are coherent with the estimated sensitivity of DHFR PCA and validate our experimental scheme (Table 11 and Fig. 1*B*; see also Fig. 10, which is published as supporting information on the PNAS web site).

Discussion

Above, we presented a general strategy to expand the sequence space explored by a fold. This approach was used to generate a massive (72 of 76 residues were perturbed) sequence perturbation of a small protein. Strikingly, this set of functionally related sequences derived from the Raf RBD closely approximates the sequence space observed in sa MSAs as shown by the entropy profile comparisons (Fig. 2*E*). Nevertheless, the inner core residues, which are among the most constrained residues by the absence of long-range covariation, display, in most cases, at least a slightly predominant selection for wt aa (Fig. 3 and Table 9). However,

comparison of our results versus studies reporting partial degeneracy of a small protein such as λ-repressor, barnase, protein-L, and ubiquitin (10, 11, 14) reveal similar behavior of core residues despite variation in the location and dispersion of degenerated positions (e.g., concomitantly on a series of core residues or on a short segment of residues), suggesting that it might be stemming more from the limited number of residues covaried in each of these studies than from the nature of sequence perturbation. The potential bias introduced by fixing the sequence around a degenerated segment could be compensated for by the lower overall destabilization penalty induced by the segmental sequence perturbation versus corandomization of core residues. Alternatively, it might indicate that the local nature of structure formation has been underestimated. Lau and Dill (3) had observed that proteins are remarkably robust to point mutation, which is the main motor of natural sequence space expansion in fold evolution, although this type of alteration occurs at a very low rate in normal conditions (10^{-10} to 10^{-11} ·bp^{-1}·replication^{-1}) (25, 26). It is also noteworthy that the most successful algorithm for predicting structure and *de novo* design, Rosetta, is based on optimizing local elements of structure on successive stretches of residues (27). In conclusion, the induction of massive sequence perturbation through a segment-by-segment approach does not reveal the complete sequence space compatible to a fold, but as reported here, it is sufficient to outline the repertory of sequence variation and constraints imposed upon it.

At present, there is no reliable method to screen for sequences capable of forming a specific target fold (28). Previously, studies have used protein–protein or protein–peptide interactions to screen libraries to select clones capable of folding into a specific structure (13, 14, 29). The residues of c-Raf RBD directly involved in binding to *h-ras* were thoroughly identified in a preceding mutagenesis study (22). However, to detect any other deviation in the experimental aa selection introduced by the experimental method, we compared the experimental data with sa and fh MSAs recovered from databases. Based on these comparisons, we proposed a signature sequence for Raf RBD (Fig. 3). Most of these consensus positions are also conserved in the sa MSAs and, thus, are consistent with the proposal that they are also key features of the ubiquitin superfold. The number of residues in the signature sequence versus Raf RBD length (30 of 76) is consistent with the average sequence identity observed between redesigned proteins and their natural counterparts as reported in ref. 5. These results suggest that the experimental strategy outlined here can produce a MSA containing significant structural information. Interestingly, Rosetta performance in structure predictions was improved by adding MSA information to its regular algorithm routine (4). As demonstrated by this example, the capacity to artificially extend sequence space explored by any (poorly populated) folds could be helpful in protein design and structure prediction.

Lockless *et al.* (30) have proposed a method to identify mechanistically coupled functionally important residue pairs. This approach necessitates large MSAs to test pairs of residues for covariation. One could easily design experiments based on our strategy to generate large RBD sets in place of two or more remote segments of the primary structure or in a mutant background either of the RBD or of its binding partner *h-ras*. Residues important for their dimerization and showing specific convergence in the experimental aa selection and SMART MSA (Fig. 1*B* and 3) could constitute a good starting point for these investigations.

Finally, the experimental data and the sa MSA reveal a two-layer assembly of the hydrophobic core of the Raf RBD and of the ubiquitin superfold, which is reminiscent of a measure of global hydrophobic core formation that is based on micelle-like models used, coincidently, to improve Rosetta performance (4). Furthermore, structural observations of proteins sharing the ubiquitin-roll topology suggest that the spatial dispositions of residues found in the inner and outer core are conserved. Interestingly, the volume of inner core residues side chain is fairly constant across this superfold, particularly in superfamilies grouped in the SCOP MSA, with an average volume of 594 ± 49 Å^3 (Table 6). Similar observations were described for data obtained by degenerating concomitantly core residues of λ-repressor and barnase (10, 11). Moreover, theoretical studies have suggested that conservation of core volume might be particularly relevant for domains smaller than 200 residues, because they usually have higher packing density than larger domains (31). Therefore, combinations of volume density dispersion at key positions in MSAs with simple graphical representation of protein structure could represent a way of unraveling unsuspected architectural or evolutionary linkages between sa and structure topologies (32, 33). Such additional geometrical constraints could be useful to improve structure prediction and protein design algorithms.

We thank J. Bonvin and J. Turnbull for circular dichroism; P. Lepage, D. Roquis, and M. Fyfe for sequencing; S. Bilodeau and T. Viet for NMR; and A. Vallée-Belisle, S. Pontier, and M. Lerch (discussions). Natural Sciences and Engineering Research Council and Canadian Institutes of Health Research (CIHR) funded this project. F.X.C.-V. was a CIHR, Program de Biologie Moléculaire, and Faculté des Etudes Supérieures scholar. S.W.M. is the Canada Research Chair in Integrative Genomics.

1. Perutz, M. F., Kendrew, J. C. & Watson, H. C. (1965) *J. Mol. Biol.* **13,** 669–678.
2. Lesk, A. M. & Chothia, C. (1980) *J. Mol. Biol.* **136,** 225–270.
3. Lau, K. F. & Dill, K. A. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 638–642.
4. Bonneau, R., Strauss, C. E. & Baker, D. (2001) *Proteins* **43,** 1–11.
5. Dantas, G., Kuhlman, B., Callender, D., Wong, M. & Baker, D. (2003) *J. Mol. Biol.* **332,** 449–460.
6. Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L. & Baker, D. (2003) *Science* **302,** 1364–1368.
7. Scalley-Kim, M. & Baker, D. (2004) *J. Mol. Biol.* **338,** 573–583.
8. Arkin, A. P. & Youvan, D. C. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 7811–7815.
9. Reidhaar-Olson, J. F. & Sauer, R. T. (1988) *Science* **241,** 53–57.
10. Lim, W. A. & Sauer, R. T. (1989) *Nature* **339,** 31–36.
11. Axe, D. D., Foster, N. W. & Fersht, A. R. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 5590–5594.
12. Finucane, M. D. & Woolfson, D. N. (1999) *Biochemistry* **38,** 11613–11623.
13. Gu, H., Kim, D. & Baker, D. (1997) *J. Mol. Biol.* **274,** 588–596.
14. Kim, D. E., Gu, H. & Baker, D. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 4982–4986.
15. Emerson, S. D., Madison, V. S., Palermo, R. E., Waugh, D. S., Scheffler, J. E., Tsao, K. L., Kiefer, S. E., Liu, S. P. & Fry, D. C. (1995) *Biochemistry* **34,** 6911–6918.
16. Vallee-Belisle, A., Turcotte, J. F. & Michnick, S. W. (2004) *Biochemistry* **43,** 8447–8458.
17. Soding, J. & Lupas, A. N. (2003) *Bioessays* **25,** 837–846.
18. Campbell-Valois, F.-X. & Michnick, S. W. (2005) *Methods and Protocols in Molecular Biology*, in press.
19. Pelletier, J. N., Campbell-Valois, F.-X. & Michnick, S. W. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 12141–12146.
20. Sander, C. & Schneider, R. (1991) *Proteins* **9,** 56–68.
21. Pelletier, J. N., Arndt, K. M., Pluckthun, A. & Michnick, S. W. (1999) *Nat. Biotechnol.* **17,** 683–690.
22. Block, C., Janknecht, R., Herrmann, C., Nassar, N. & Wittinghofer, A. (1996) *Nat. Struct. Biol.* **3,** 244–251.
23. Fridman, M., Maruta, H., Gonez, J., Walker, F., Treutlein, H., Zeng, J. & Burgess, A. (2000) *J. Biol. Chem.* **275,** 30363–30371.
24. Fridman, M., Walker, F., Catimel, B., Domagala, T., Nice, E. & Burgess, A. (2000) *Biochemistry* **39,** 15603–15611.
25. Grishin, N. V. (2001) *J. Struct. Biol.* **134,** 167–185.
26. Drake, J. W., Charlesworth, B., Charlesworth, D. & Crow, J. F. (1998) *Genetics* **148,** 1667–1686.
27. Simons, K. T., Strauss, C. & Baker, D. (2001) *J. Mol. Biol.* **306,** 1191–1199.
28. Waldo, G. S. (2003) *Curr. Opin. Chem. Biol.* **7,** 33–38.
29. Riddle, D. S., Santiago, J. V., Bray-Hall, S. T., Doshi, N., Grantcharova, V. P., Yi, Q. & Baker, D. (1997) *Nat. Struct. Biol.* **4,** 805–809.
30. Lockless, S. W. & Ranganathan, R. (1999) *Science* **286,** 295–299.
31. Liang, J. & Dill, K. A. (2001) *Biophys. J.* **81,** 751–766.
32. Kannan, N. & Vishveshwara, S. (1999) *J. Mol. Biol.* **292,** 441–464.
33. Lindorff-Larsen, K., Rogen, P., Paci, E., Vendruscolo, M. & Dobson, C. M. (2005) *Trends Biochem. Sci.* **30,** 13–19.

Table 2. Sequence of clones selected through our experimental strategy and used for statistics Eqs. 1 and 2

S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13
SNTIRVF	**LPNKQ***	**RTVVN**	**VRNGMS**	**LHDCLM**	**K*ALKVR***	**GLQPEC**	**CAVFRL**	**LHEHKGK**	**KARLDW***	**NTDAAS**	**LIGE**	**ELQVDF**
VTLVKVF	CPDGQ*	ASIKH	ARRGKS	VETHLH†	K*SMEHR*	SVPQNA	RLTVSR‡	SLDCFLT	VMNMPW*	GMCAAK	LCRE	DIQLEV
AEFFTVF	IIDGQ*†	RTMLM	IQWGTS‡	VENMLK	K*RMELR*	SFMQGH	TLTVWR	TLLCRVT	HDNLPW*	AMAAAF	STVE	HIHLEI
TNLMAVH	LQDGQ*	NSTMH	HKIGTG	VEKALR	K*SMVRR*	ALNQNE	KLGVRA	TVGCCSV	VAILDW*‡‖	SLDAGD	LTSE	HLFLEL‡
PMTMTVH	LADGQ*	WSIQR‡‖	PSKGAM	VERALQ	K*SMVLR*	EMLLND	KLVIAR‡§‖	DCIKRTN	MARLDW*	NRDAGK‡	LTGG	ELYTEI
MSWMPVY	IGQFQ*	SSYMR	PKKGLM	VKTTLF	K*SMGKR*	RVCLNE	MDRPWL	HTILRPT	TNRVDW*	FSPARD	LRGL	AIKIEF
RSLSCVH	IGLAQ*	QSTAV	PKCGLT	VSMILE	K*SMGLR*	KMCANR	MTRPVL	STHEGPV	RSALGW*	HSMAKD	LSGL	LIAVES
RGLMCIW	IGECQ*	RSTAS	VDRPRD	VRMILA	K*TMGHR*	KMDFVT	MKKPRL	ADTLLVY	MSALTW*	ECSARD	LRGK	KLVISP
RRLCFVM	IGSSQ*	KSTVV	MDRWAL	VHMVLL	K*TMYHR*	DCKAEQ	IDRLRL	TPTLHSD	NSRLLW*	ARVARG	LGSS	KLVIDY
ATCLFID	IGTSQ*	RSTVW	PSRPTS	ILDLVM‡	K*TMAMR*	NCPSEC	MSRIRL	SCTSKRD	GIRLQW*	HYSASV†	LPQS	GLVIEP
LMCMFVD	ILESQ*†	ISSVT	ANLEMT	ISDLMI	K*PMVSR*	QCDRET	LRGIRL	SHTYFRP	SRDLVW*	EMQASV	LRCS	SIVLQP
LLYLVVH‡	ISGDQ*	MSVVF	AMLGLT	IQQAVD	K*PMKQR*	SLSKEY	LICLRL	PDQSATG‡‖	LRLDVW*	NAFAST	LRHC	NLTIVS
LVLLEVL	ILADQ*	HSYCY	ARLSET	IRIHLN	K*PMDTR*	SYIPEY	GLTATL	EDQERIG	TRCLSW*	CVSNSA†	LSVN	NITYWQ
AMSLCCL	IQRTQ*	ISCCS	AGPVHT	IVDHLR	K*KMFSR*	SMSSES‡§¶	SLEARL	GPGQLTN	CRCLDW*	DVSCST	LADN	SLWCKQ
LTHLKCL	LGVHQ*	YALIA‡	CVLTDT	ISICLS	K*KMVTR*	GTSSQA	NMVIRI	TYGQGMG	PPLLMW*	QQDASA	LGVT	RIYVKE‡
LPKLNVK‡	LGAEQ*	YSRII	CGNSGT	LRWNLS	K*VMWRR*	GEDRQA	NMVLRK	MEPHIMT	PGWLEW*	NCDSSA	LGVL	RISVCT
TPRLKVG	LGNAQ*	LSTIA	GQCAMT	LKDNLS	K*AMCRR*	RVDSQQ	NMKVFM	GESHSTG	IPELYW*	PGCRYV	LNPQ	SIEVLE
CMRLYVK	LGHAQ*	MIALD	ACCCMT	LGVELS‡	K*KLSER*‡§¶	WLDRQF	MKVRRR	PEGSGSN	LDELHW*	PGWCME	ITPT	HIIVSH
SSRLLVL	VGHYQ*	HCRLS	SPRRET	LVKNLQ	K*KLTSR*	RISWRY	MKVQRS	PTKSKAE	LMMLNW*	GGIRSH	ISQT	HIVVHS
TVPLNVS	LGSSQ*	FSWLG	VPLPQT	LSKNLE	K*KLFMR*	RIGIGD	MKVLQM	PYRESSE	WNQLGW*	KLQRPA	IRGT‡	KICVTC
MMPLWVY	LGSTQ*	QCCAQ	VNEVVR	LYDCLM	K*KLEMR*	NIACVS	MVVERV	PYEVMSF	WQSLKW*	GQDDFK	KTTT	MVLVYE
SFNLAVK‡§¶‖	LPRLQ*‡§¶‖	QINPM	VNTPVG	LYICLI	K*KLCRR*	RIPRAS	MVVCRW	PQPNSVD	RNLLAW*	GSTPEK	SQTT	DGLVQE‡
TFMTRVW	LPHLQ*	DGNVY	VAKPLP	LQRCLQ	K*TLIRR*	DRMPSD	MAVWGG	GWLCMRR	RCALAW*	GMDGSR	LPTT	KVFIVE

Table 2 (continued).

S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13
SNTIRVF	**LPNKQ***	**RTVVN**	**VRNGMS**	**LHDCLM**	**K*ALKVR***	**GLQPEC**	**CAVFRL**	**LHEHKGK**	**KARLDW***	**NTDAAS**	**LIGE**	**ELQVDF**
RWALAVA	LPTMQ*	LTTVN	AYGKRP	LKRCLE	K*TLEKR*	TIIPTD	LRVNRT	GEGENRR	RVGLAW*	QMDVSQ	MPTG	TVDVVN
LWPLPVA	LPGHQ*‡§¶‖	GTRVT‡§¶‖	VYGKWP	LCSSLD	K*GLQKR*	DIRRKD	MGVNRT	ANAMPRE	RHGLTW*	GADGRE	RMLD	KVVVNT
PWVDLDA‡§¶	LPFVQ*	WTMVR	VKWKRE	LKQSLL‡	K*ALSKR*	VLTVQD†	ISVWRT	HNRRHEA	HVTRAW*	GHDVRT	MDLT†	EVAVDT
KSRVAVA	LPLVQ*	CTQSR	VPASSL	LYSTVL	K*RLLMR*	HLREQD	VEVFML	MNWESEG	HVKIAW*	HDDPND‡	NART	CITVDN
QSGIAVA	LPLTQ*	LTQMR	VFASQR	LYTTVY	K*RLNIR*	KLMGAD	TEVFRR	ARGRRLA	RVSVAW*	TFDTSD	LTRT	SLSVDN
GSTVIVA	LPMGQ*	LTSAL	VPAQIP	LMATLL	K*RLWNR*	MLAYAD	TSALQI	ASRARLQ	RIPTLW*	GSPCSQ	LSCT	RLMVSD‡§¶
TSCLRVA	LPSWQ*	LTSKC	VQAAES	LATTLH	K*CLFNR*	KLQRGD	TSITRI	RNGGRAD	RISTEW*	ASPMSN	VSRY	RLELLD
RDWFPNN	LPSAQ*	LTSSS	VACASS	LQTTLQ	K*RLASR*	SLQRAN	ASITLV	LNGLSKA	RQSTYW*	HSRCES	VGGS	SLAVHD
RYWVVAQ	LPSSQ*	LTIVS	VLHHPV	LRTLIG	K*RLALR*	KLSASH	RSDLIV	LIGSESA	RVRQYW*	ACEGSS	VGGG	SLKVVH
RMTVAIN	LPSNQ*	LTETI	VLGAIE	LRTLMK	K*VLAFR*	DLAVSH	MSDLRV	IDNFQSA	RMTATW*	NSEGYA	MGGS	NLWVRS
CRSLLAN‡	LPHNQ*†	WTQQL	VMRSIQ	LRTVLL	K*ALCYR*	SLSGST	IEQLTV	RSSVIPA	RMMAQW*	KVDSTR	ERGS†	YLRVDS
KMLLLAW	LPENQ*	ASTQL	TARAIA	LLGLLK	K*RLHYR*	TLKASD	LTTTRV	DFKVAFA	RMAALW*	DWDSVS	QDGF	QLYVSS
LPCLIAR	LTDNQ*	KTTQL	CGSNLR	LKRVCS	K*RLVHR*‡	ALRHSR	TTVYRV	NSTTNTA	EVTLPW*	MTPSTS	NVGW	RLCVKS
SRKLMAR	LTDHQ*	KQPKL‡	IGSRGM	LLDVSR	K*RLVSR*	DLGHDG	MTVKFK	EDGDGVP‡	MVTLPW*	DRESSE	VQDH	SLVVTS
RELLAAS	LPDHQ*	VTTKL†	VSSRAR	LNHILG	K*KLVWR*‡§¶	NLDKLS	MTIRSK	EILPGSI	IVTPGW*	CAPSWE	IQGH	TLTVTA
AAVVSAV	LPDNQ*	VTIGL	VGKAAR	INNILE	K*SLVTR*	NLDLKE	MFAIKA	IRWRGRV	EVLAEW*	DAPWSS	GQGC	SLTVRA
GCYVYAS	MPETQ*	VTTRI	TLHLRR	LQNIIV†	K*SLGSR*	NMFFKG	MEARMC	SRARGKA	LPATPW*	DQPIQF	HMGH	SLTVEV
HSATVAH	MPECQ*	STIMG	DRSLTR	LYEIIN	K*SLGNR*	NMKLKN	MVPRAA	NRKRPMP	LVAVTW*	SCPGQD	HLGA	TLSVLL
WSAVSFE	LPECQ*	KTIRM	LHGRKH	LRDELA	K*SLWSR*	NLQSTQ	MTELVA	RFKRGSK	ATQTDW*	QTDTEG	YLGL	KLVVCL
FAIIKFQ	LLETQ*	MTMFV	LGRHYH	VRDGLA	K*SLKNR*	NYRACH	MYIQRC	SIEGLNK	WTPADW*	NCDPEG	MRQR	KLSVEA
FAPITAI	LPESQ*	HTYRV	PAPGIV	LRKILA	K*SLKMR*	AASVEQ	MEIHRS	MSVRSNY	LTHHPW*	RHEMEG	MRGH	DVSVRM
FQALTAM	LPDRQ*	CTLRH	ASPGRV	LRDIMV‡	K*SCKIR*	AITVPQ	MYLTRR	SSVRINK	STHTGW*	YCNVDD	SRYI	DRMVRV

Table 2 (continued).

S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13
SNTIRVF	**LPNKQ***	**RTVVN**	**VRNGMS**	**LHDCLM**	**K*ALKVR***	**GLQPEC**	**CAVFRL**	**LHEHKGK**	**KARLDW***	**NTDAAS**	**LIGE**	**ELQVDF**
SCCFTAM	LPYRQ*	RTLQS	ARPRIL	LRDILV	K*SLMIR*	GWSAPQ	LGLPTR	TSRPSVS	MAAIKW*‡	DCNVDM	CLSS	DLKVQM
TRLIRIS	LPERQ*	YTLTC	AWSPIK	LREILR	K*SVMRR*	GVQGPD	LKLPFQ	HSRCTDK	IASIAW*	DSNSED	CLQS	CLEITM
YRPIVAW	LANNQ*	AVLNR	APILGS	LEYILR	K*SALRR*	GWRRGR	LGLHSV	RPRIKFV	DDAVAW*	RTNSEA	CQES	LCEVTK
RSVITVS	LCNGQ*	ATTLK	PMSSSS	LRASLR	K*SLSRR*	GSGVGR	LGLVVE	SPRSDSQ	WAAVDW*	TISCCT	CLTE	DMFVDA
GQVITCN	MPNSQ*	YTVLR	FPPESS	VRGCME	K*SLRHR*	GYPWGG	LVLTMG	NQYPRGV	IAGVSW*	RINCDT	CLGE	TLMLTR
AMNISVS	LPNSQ*	ITARQ	LLPGVT	LRKMME	K*SLRRR*	GTHVNK	LVLASM	EKRPKGA	QAVVNW*	RTDCIT	CCQL	TSCHDL‡§¶
DDMIACV	CPNMQ*	ITCTM	ILPSVS	LVQIME	K*SLHVR*	QRVVMV	LTPPEM	AGYNEPL	RAYKRW*	TTDCTP	CVSF	RSNPQL
MTALEFK	LPNDQ*	ATVTQ	PLANVP	LEVLMS	K*SLLVR*	LGVPGY	LAKPET	AETSETL	SATLRW*	SMSTRT	CAST	PSHNTA†
DTVLFFR	LPNFQ*	FSFRR	RSATAA	LSVSMG	K*HLMWR*	GYVLTA†	LGCYGP	GVNLESC	SISITW*	SMTCDD	AVNS	PVNMMS
RTNLTFQ	LPKEQ*	FTMRR	RWASMS	LGEAMC	K*GLRLR*	GFVSKS	LTPYIN	KVMLELL	SLEITW*	GLKGTS	IVNN	TFNGIP
GWMLMFD	LPWEQ*	FAVRM	EFVTLL	VAQLCQ	K*CCRLR*	GMDTNE	LEILPL‡	NKGNEFS	SFETEW*	TMKATP	CRNQ	SKFWMV
TGMLPFS	LPNEQ*	FMVTS	TTWTCY	VAESLS	K*PLGLR*	GLLTQE	LAIALL	GCYGTKS	SWLVGW*	STATTD	VNCA	IMQITV
HHQLEFA	MPDQQ*	VAQSI‡	KTTPPL†	AAEMLT	K*ACVQR*	GLALRE‡	LKINFR	YGTRVGS	SCMVRW*	GTAATT	TGAA	LLYIEI
NSQLEIF	IPFQQ*	VSQTF	FTENMK‡	LAETMT	K*ALQQR*‡‖	GLRLAS	LTIWFG	EGSKVDM	GPSASW*	KATATS	HRDA	PNRMIV
ANYLTTR	FNSSQ*	NAQVV	NIMSRR	LEAQLS	K*AVYRR*	GVESDL	LMATLQ	ERATYDS	IGSASW*	QNTATS	LVDA	MLDVRQ
AATLNIR	FNDGQ*	SSQVV	EIEDRA	VMGCLL	K*AVVSR*	GVLVDC	LSAAIK	EVDVLDE	GLSLSW*	KQCATL	LMDR	RLTCLS
PCTLSCS	FGDAQ*	RHVST	SWYQRS	LFNSIH	K*AFRSR*	GLERDS	LGITYK	ELEKHPQ	NLSLSW*	KTTAKE	QTDC	RLEITQ
NRTLICY		VHTTT	KPPNRC	LKDTVH	K*AMRVR*	QLDMHA	MVVYSR	EATKLML	GYTLSW*	FETVTG	VGDC	TEISVE
PPTLRIM		RSMLR	GRIARK†	LSRMVE	K*AAKLR*	GLTGHA	LVVFSR	NLSPLCR	LGRVSW*	SVTAME	RIPG	RLLVNN
VCRCRIK		RKMQA	QRWYAC	LQVLMC	K*ALKVR*	GLTAVT	LYVRSR	MLADCCS	LITVSW*	SVQAMS	ICGH	ACDVEQ
GCSLAIA		VTERT	HPLPSC	VSATLA	K*SFLTR*	GLQPEC	LRVRYF	QLLADIQ	KYPVSW*	SVEAEH	YAGT	
		MSAIP	ISTPCK	LINIMV	K*KFLLR*	ELNHRE‡	LRGVYM	QAWVDVS	VLMVSW*	SDGASH	QDGF	

Table 2 (continued).

S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13
		RTVVN	**VRNGMS**	**LHDCLM**	**K*ALKVR***	**GLQPEC**	**CAVFRL**	**LHEHKGK**	**KARLDW***	**NTDAAS**	**LIGE**	
		RHVMQ	CSPSGR	LEESIM	K*KFACR*	SCGIED	LRVARN	QSNKCET	HLVQSW*	AKGNMC	QRGW	
		STRLS	ALTEGN	LVALLR	K*VFSVR*		LRVKRK	QMNPGAL	TASMPW*	QLPAVR	LLNN	
		RSIMQ	ALLGHK	LEENLH	K*VFKHR*		LTVRNE	LYPIAES	TASPCW*	HQPANR	LLPA	
		RTLLP	AGPEMC	LQDFIG	K*QFVNR*		LTVRCS	WANQATF	TRSVQW*	EKPAER		
			FQPEQT	VINCVK	K*PFLNR*		LDVQME	GEQYPNL	TRSFSW*	RMPAGS		
			PGPRLS	VKISLM	K*PILVR*		LMVSKA	RAECKGA	NGESQW*[†]	SRNAAS		
			VSPRSS		K*PIRSR*		LAISRR	ELDIWEK	NEPMQW*			
			LKPYNT		K*PIRMR*		TAVSVR	IDKSEIV	KLSEQW*			
			IKPLNS		K*HINCR*		LSIGAR		KLEAQW*			
			IPPKSK		K*CINHR*		MSISAR		TLESGW*			
			ARRMQT		K*SINLR*		TSIRYR		TLLLGW*			
			VSVAQT		K*GIMLR*		LSLSYR		VESVQW*			
			VRKNER		K*KIFVR*		TCLNHI		HESVLW*			
			VRHPET		K*KIWGR*		TCIYVE[†]		EEKVVW*			
			ATPRST		K*RIGFR*		TWIHSM		MEKLHW*			
			ACVRST				TCVSLD		GESYRW*			
			CTCTST				TYVTLK					
			VAEYST				DKLGIW[‡§]					
			LLPAVR				ADLMIW					
			ALAALT				TRMGLW					
			VLSRVT				TRVGWR					

Table 2 (continued).

S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13
			VRNGMS				**CAVFRL**					
			GFSGVT				CKVGWG					
			PSARVT				GIMIWH[†]					
			IVQGGT				SRYTVS					
			IVKRYT				MLYGLY					
			IRPGLT									
			CHSKLS									
			CCGTLS									
			TTTALS									
			IHRAES									
			ARAPRS									
			ASGPGS									
			ATGVSS									
			PTRSSS									
			PASSYS									
			PDNQQS									
			ISSPVS									
			INSSMP									
			VCTETS									
			VTTGKK									
			VHMSIR									
			VHEATG									

Table 2 (continued).

	S4
	VRNGMS
	VLLSRP
	VLSSRV
	VLARAP
	VEPLKP
	VMPLAS
	VPRTCA
	VPRPMN
	PHRNCN
	EHLTPK

*Residues not degenerated in the experiments.

†Clones that were not purified with sufficient yield.

‡Clones that were successfully purified.

§Clones for which CD spectra were done. Some of the spectra were presented (Fig. 5*A*).

¶Clones for which chevron curves are presented in Fig. 5*B*.

||Clone for which affinity curves were done (Table 11).

Table 3. Sequences of clones isolated through screening of three alternative libraries

S2b[†]	S6b[‡]	S8b[§]
FNGSG	DRLANR	LSIRR*R
FSGQG	KKLVVR	MSIRR*V
FSDLG	HSLRTR	MLISR*P
FYQLG	RPISMR	LVLKR*K
FPDYA	APMSHR	LVLFR*Y
MPHEQ	MPMTRR	LTHPR*Y
LPMEQ	RAMELR	LRVER*R
LPQHQ	STIKRR	TFVKR*K
LNGGQ		TWISR*R
LGSTG		TMLSR*R
LGSTM		TYIMR*T
LGTTQ		CIIHR*K
LADST		MWVTR*L
		KLQIR*L
		VMIYR*L

†S2b is identical to the main experimental library S2 except that Q66 is also degenerated (Table 1 and 2).

‡S6b is identical to the main experimental library S6 except that K84 and R89 are also degenerated (Table 1 and 2).

§S8b is identical to the main experimental library S8 except that R100 is kept constant (Table 1 and 2).

Table 4. Alignments of functional homologues (MSA S1) and structural analogues (MSA S2 and S3; note that wt Raf RBD sequence, 1RFA, is highlighted with bold lettering)

MSA S1. Functional homologues of Raf type RBD (SMART)

smart|RBD-sptrembl|Q8IN00|Q8IN	SLCRVIL-TDGATTIVQTRPGETVGELVERLLEKRNLVYPYYDIVF--QG---STK-SIDVQQPSQILAGKEVVIERR
smart|RBD-ENSANGP00000001206/3	TLCRVIL-SNGATTVVQTRSNETIKELVERLLEKRGIVYNAYEAFL--AG---STK-PLDLDGPSVSLAGKEVNIDQR
smart|RBD-ENSMUSP00000021945/2	KYCCVYL-PDGTASLALARPGLTIRDMLAGICEKRGLSLPDIKVYL--VG---NEQKALVLDQDCTVLADQEVRLENR
smart|RBD-ENSMUSP00000030984/3	KHCCVHL-PDGTSCVVAVKSGFSIKEILSGLCERHGINGAAVDLFL--VG---GDK-PLVLHQDSSILATRDLRLEKR
smart|RBD-SINFRUP00000158915/2	RQCRVML-PEG-SCSSILRPGSFIREVLQDLCQSIGVNIAAVDLFL--VG---GEK-PLVLDQDCMTLCSRDLRLEKR
smart|RBD-sptrembl|Q13878|Q138	PIVRVFL-PNKQRTVVPARCGVTVRDSLKKALMMRGLIPECCAVYR-I------------------------------
smart|RBD-sptremblnew|BAB32131	GTVKVYL-PNKQRTVVTVREGMSVYDSLDKALKVRGLNQDCCVVYR-LIK---GRKTVTAWDTAIAPLDGEELIVEVL
smart|RBD-sptrembl|Q98TC3|Q98T	STIRVYL-PNQQRTVVNVRPGMTLHNCLIKALKVRGLQPQCCAVFR-LHPGQRSKKLRMDWNTDSTSLIGQELLVEVL
smart|RBD-swissprot|P09560|KRA	STMRVYL-PNKQRTVVNVRSGMSLHDCLMKSLKVRGLQPECCAVFR-LIQDPKGK-LRLDWNTDAMSLVGAELQVDFL
smart|RBD-sptrembl|Q8I086|Q8I0	ILLRAHL-PNQQRTSVEVISGVRLCDALMKALKLRQLTPDMCEVST-THSG----RHIIPWHTDIGTLHVEEIFVRLL
smart|RBD-ENSANGP00000003633/1	MLLRAFL-PNQQRTSVQVIPGMRLKDALAKALKRRNLTCEFCEVTA-GNS-----NYPIPWETDVSALNCDEVFVRIL
smart|RBD-sptrembl|Q8MXT8|Q8MX	KMIMVHL-PFDQHSRVEVRPGETARDAISKLLKKRNITPQLCHVNASSDPKQESIELSLTMEEIASRLPGNELWVHSE
smart|RBD-sptrembl|Q9GT28|Q9GT	SLILLHL-PFNQHSKVEEKPGILARDAIAKILEKRAILPQMCRVCVGGDPSSPRTDLSMDLETLSGALEKKELWVHSA
smart|RBD-ENSANGP00000002581/1	KSYKVAL-PENTFATVYLREGMSVEEFLASACSRKNLNPMEHFVRVKKRRDM--EDHNYFVPHR------NDL-IETY
smart|RBD-ENSMUSP00000002588/7	TPSWFCL-PNNQPALTVVRPGDTARDTLELICKTHQLDHSAHYLRLKFLME---NRVQFYIPQP-----EEDI-YELL
smart|RBD-SINFRUP00000133952/7	TPSWVCL-PNDQPVLTIIKPGESALCVLESICKAHYLDPTRHYLRLKFLME---SQVKIYIPKP-----DEDV-CDLV
smart|RBD-ENSMUSP00000024562/8	VQTYVHF-QDNEGVTVTIKPEHRVEDVLALVCKMRQLEPTHYGLQLRKVVD---KSVEWCVPALYEYM-QEQASYDEI
smart|RBD-ENSP00000275245/186-	IQTYVHF-QDNHGVTVGIKPEHRVEDILTLACKMRQLEPSHYGLQLRKLVD---DNVEYCIPAPYEYM-QQQV-YDEI
smart|RBD-sptrembl|Q8IN00|Q8#I	VAFKLDL-PDPKVISVKSKPKKQLHEVIRPILSKYNYKMEQVQVIM--RDT----QVPIDLNQPVTMADGQRLRIVMV
smart|RBD-ENSANGP00000001206#/	VVFKLNL-PNRKMISVKSKAAKPLADVLRPILHKYNYELDEMKVVV--HSTV---DVCLDMTQPVTTVDGLCLYIRSA
smart|RBD-ENSMUSP00000021945#/	-TFQLELVGLERVVRISAKPTKRLQEALQPILAKHGLSLDQVVLHR--PGE----KQPMDLENPVSSVASQTLVLDTP
smart|RBD-ENSMUSP00000030984#/	-LFRLDLVPINRSVGLKAKPTKPVTEVLRPVVAKYGLDLGSLLVRL--SGE----KEPLDLGAPISSLDGQRVILEER

Table 4 (continued).

MSA S2. Structural analogues classified in the ubiquitin superfamily and 4 closely related super-families (SCOP)	
1RFA	**NTIRVFLPNKQRTVVNVRNGMSLHDCLMKALKVRGLQPECCAVFRLKARLDWNTDAASLIGEELQVDF**
1UBI	MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYKESTLHLVL
1A5R	IKLKVIGQDSSEIHFKVK--MHLKKLKESYCQRQGVPMNSLRFLFEGQRIADNHT-PKEEEDVIEVYQ
1MG8a	MIVFVRFNSSYGFPVEVDSDTSILQLKEVVAKRQGVPADQLRVIFAGKELPNHLTVQNCQQSIVHIVQ
1VCBa	VFLMIR-RHKTTIFTDAKESSTVFELKRIVEGILKRPPDEQRLYKDDQLLDDGKTLGECAPATVGLAF
1M94a	IEVVVNDRLGKKVRVKCLAEDSVGDFKKVLSLQIGTQPNKIVLQKGGSVLKDHISLEDYDQTNLELYY
1J8Ca	IKVTVKTP-KEKEEFAVPENSSVQQFKEAISKRFKSQTDQLVLIFAGKILKDQDTLIQHDGLTVHLVI
1H8Ca	SKLRIRTPSGEFLERRFLASNKLQIVFDFVASK-GFPWDEYKLLSTRRDTDPNKSLLEVPQETLFLEA
1JRUa	TNIQIRLADGGRLVQKFNHSHRISDIRLFIVDA-AM--AATSFVLM-KELADQ-TLKEAAVIVQRLT-
1EO6a	VPVIVEKV-----KYLVPSDITVAQFMWIIRKRIQLPSEKIFLFVDKTVPQSSLTMGQLGFLYVAYSG
1EF1a	ISVRVTTD--AELEFAIQPNTTGKQLFDQVVKTIGLEVWFFGLQYQSTWLKLNKKVTAQSPLLFKFRA
1GG3a	MHCKVSLLDDTVYECVVEKHAKQDLLKRVCEHLLNLDYFGLAIWDNKTWLDSAKEIKKQ-PWNFTFNV
1LFDa	CIIRVSLDV--YKSILVTSQDKAPTVIRKAMDKHNLEPEDYELLQIKLKIPENANVFYA-NYDFILKK
1E8Xa	VFIVIHRS-TTSQTIKVSADDTPGTILQSFFTKMA---RDFVLRVRDEYLVGETPIKNF-EIHLVLDT
1K8Rb	CILRFIACNGQTRAVQSRG--DYQKTLAIALKKFSLDASKFIVCVSIKLITEE-------D-RLIIVP
1L7Ya	VTFKITLSD-PFKVLSVPESTPFTAVLKFAAEEFKVPAATSAIITNGVGVNPAQPAGNIFGSELRLIP
1D4Ba	RPFRVCDHKRIRKGLTAA---QLAKALETL-----L-NGVLTLVLE-TAVDSEDFFQLL-DTCLMVLQ
1C9Fa	KCVKLRALHSCKFGVAAR---SLLRKGCVRF---QLPMPGSRLCLDGTEVTDDFP---LNDAELLLLT
1F2Ri	KPCLLRRNHSDQHGVAAS---SLRSKACELL---AI----ITLVLAGTIVDDDY-FLCLSNTKFVALA
1IP9a	TKIKFYYK-DDIFALMLKGDTTLRSKIAPRI---DTD---FKLQTKSEEIKTDQ-VSNIIALKISVHD
1Q1Oa	ILFRISY---EIFTLLVEKVWNLIMAINSKISNT------IKIKYQFVVLGSDD-WNVAKFLNIRLY-
1FMAd	IMIKVLFF---ATEVAADFP-TVEALRQHMAAQSDLEDGKLLAAVNQTLVSFDHPL---DGDEVAFFP
1F0Za	MQILF--N-D--QAMQCAAGQTVHELLEQL----DQRQAGAALAINQQIVQWAHIV---DGDQILLFQ
1JSBa	MKFTVITD-DGKKILESGAPRRIKDVLGEL----EIPIETVVVKKNGQIVIDEEEI---DGDIIEVIR
1QF6a	PVITL--P-D-GSQRHYDHAVSPMDVALDI----PL-----AGRVNGELVDACDL---INDAQLSIIT
1JALa	QTYFTAGV-KEVRAWTVSVGATAPKAAAVIH---TF---I-RAEVIWRLEGKDYIV---DGDVMHFR-
1MG4	KKVRFYRNG-FGIVYAISPDFSFEALLADLTRTLNLP----TIYTILKKIS---SLDQLEGESYVCGS

Table 4 (continued).

MSA S3. Alignment of structural analogues (SCOP-FSSP)

```
1RFA    NTIRVFLPNKQRTVVNVRNGMSLHDCLMKALKVRGLQPECCAVFRLKARLDWNTDAASLIGEELQVDF
1UBI    MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYKESTLHLVL
1A5R    IKLKVIGQDSSEIHFKVK--MHLKKLKESYCQRQGVPMNSLRFLFEGQRIADNHT-PKEEEDVIEVYQ
1MG8a   MIVFVRFNSSYGFPVEVDSDTSILQLKEVVAKRQGVPADQLRVIFAGKELPNHLTVQNCQQSIVHIVQ
1VCBa   VFLMIR-RHKTTIFTDAKESSTVFELKRIVEGILKRPPDEQRLYKDDQLLDDGKTLGECAPATVGLAF
1M94a   IEVVVNDRLGKKVRVKCLAEDSVGDFKKVLSLQIGTQPNKIVLQKGGSVLKDHISLEDYDQTNLELYY
1J8Ca   IKVTVKTP-KEKEEFAVPENSSVQQFKEAISKRFKSQTDQLVLIFAGKILKDQDTLIQHDGLTVHLVI
1H8Ca   SKLRIRTPSGEFLERRFLASNKLQIVFDFVASK-GFPWDEYKLLSTRRDTDPNKSLLEVPQETLFLEA
1JRUa   TNIQIRLADGGRLVQKFNHSHRISDIRLFIVDA-AM--AATSFVLM-KELADQ-TLKEAAVIVQRLT-
1EO6a   VPVIVEKV-----KYLVPSDITVAQFMWIIRKRIQLPSEKIFLFVDKTVPQSSLTMGQLGFLYVAYSG
1EF1a   ISVRVTTD--AELEFAIQPNTTGKQLFDQVVKTIGLEVWFFGLQYQSTWLKLNKKVTAQSPLLFKFRA
1GG3a   MHCKVSLLDDTVYECVVEKHAKQDLLKRVCEHLLNLDYFGLAIWDNKTWLDSAKEIKKQ-PWNFTFNV
1LFDa   CIIRVSLDV--YKSILVTSQDKAPTVIRKAMDKHNLEPEDYELLQIKLKIPENANVFYA-NYDFILKK
1E8Xa   VFIVIHRS-TTSQTIKVSADDTPGTILQSFFTKMA---RDFVLRVRDEYLVGETPIKNF-EIHLVLDT
1K8Rb   CILRFIACNGQTRAVQSRG--DYQKTLAIALKKFSLDASKFIVCVSIKLITEE-------D-RLIIVP
1L7Ya   VTFKITLSD-PFKVLSVPESTPFTAVLKFAAEEFKVPAATSAIITNGVGVNPAQPAGNIFGSELRLIP
1D4Ba   RPFRVCDHKRIRKGLTAA---QLAKALETL-----L-NGVLTLVLE-TAVDSEDFFQLL-DTCLMVLQ
1C9Fa   KCVKLRALHSCKFGVAAR---SLLRKGCVRF---QLPMPGSRLCLDGTEVTDDFP---LNDAELLLLT
1F2Ri   KPCLLRRNHSDQHGVAAS---SLRSKACELL---AI----ITLVLAGTIVDDDY-FLCLSNTKFVALA
1IP9a   TKIKFYYK-DDIFALMLKGDTTLRSKIAPRI---DTD---FKLQTKSEEIKTDQ-VSNIIALKISVHD
1Q1Oa   ILFRISY---EIFTLLVEKVWNLIMAINSKISNT------IKIKYQFVVLGSDD-WNVAKFLNIRLY-
1FMAd   IMIKVLFF---ATEVAADFP-TVEALRQHMAAQSDLEDGKLLAAVNQTLVSFDHPL---DGDEVAFFP
1F0Za   MQILF--N-D--QAMQCAAGQTVHELLEQL----DQRQAGAALAINQQIVQWAHIV---DGDQILLFQ
1JSBa   MKFTVITD-DGKKILESGAPRRIKDVLGEL----EIPIETVVVKKNGQIVIDEEEI---DGDIIEVIR
1QF6a   PVITL--P-D-GSQRHYDHAVSPMDVALDI----PL-----AGRVNGELVDACDL---INDAQLSIIT
1JALa   QTYFTAGV-KEVRAWTVSVGATAPKAAAVIH---TF---I-RAEVIWRLEGKDYIV---DGDVMHFR-
1FRRa   YKTVLKTPSG-EFTLDVPEGTTILDAAEEA----GY---SCLGKVVEGFVLTIAIP---SDLVIETHK
1I7Ha   PKIVIL-P---GAVLEANSGETILDAALRN----GI---TCHCIVR-ESRLSQARV---EDLVVEIPR
1AYFa   ITVHFINRDGETLTTKGKIGDSLLDVVVQN----NLD--TCHLIFE-RSRLGQICLKAMDNMTVRVP-
1PUT    SKVVYVSHDGTRRQLDVADGVS----LMQAAVSNGI---TCHVYVN-NSRLCQIIMPELDGIVVDVPD
1L5Pa   GTITAVKG-GVKKQLKFEDDQTLFTVLTEA----GL---KCICKHV-NARLAAITLGENDGAVFEL--
1E0Za   PTVEYLNY-----TMEVAEGEYILEAAEAQ----GY---NCASIVKKDVRLTIGSP---DEVKIVYNA
1FEHa   KTIII--N-G--VQFNTDEDTTILKFARDN----NI---ICTVEVE--LVTADTLI---DGMIINTNK
1HLRa   IQKVITVN-GIEQNLFVDAEALLSDVLRQQL---GL---ACSVILDGKVVRAVTKMKRVDGAQITTIE
```

Table 4 (continued).

MSA S3 (continued)	
1RFA	**NTIRVFLPNKQRTVVNVRNGMSLHDCLMKALKVRGLQPECCAVFRLKARLDWNTDAASLIGEELQVDF**
1FO4a	DELVFFVN-GKKVVEKADPETTLLAYLRRKL---GL---ACTVMLSHFSANALAPICTLHHVAVTTVE
1JROa	MEIAFLLN-GETRRVRIEDPQSLLELLRAE----GL---ACTVMIRSRAVNALMMLPQIAGKALRTIE
1FFVa	KIITVNVN-GKAQEKAVEPRTLLIHFLREEL---NL---ACTVDIDGRSVKSTHLAVQCDGSEVLTVE
2PIA	TPFTVRLRSGTSFEI--PANRS----ILEVLRDANV----TALC--TQIMVVSRA-----S-ELVLDL
1NENb	MRLEFSIY--QDYTLEADEGMMLLDALIQLKEKDSL---GLNM--NGKNGLAITPISALPGKKIVIRP
1KF6b	KNLKIEV---AFYEVPYDATTSLLDALGYIKDNLDL---GMMV--NNVPKLAKTFLRDYTD-GMKVEA
1QLAb	RMLTIRVF-FQEYKIEEAPSMTIFIVLNMIRTYDDL---GMMI--NGRPSLARTLTKDFEDGVITLLL
1JQ4a	HTITAVTEDGESLRFECRSDED----VITAALRQNI---TCKALCSLVLLCRTYP-----TDEIELPY
1KRHa	HQVALQFEDGVTRFICIAQGETLSDAAYRQ----QI---TCRAFCE-GYVLAQCRP----DAVFQIQA
1C78a	PYLMVNV--PHYVEFPIKPGLTIEYYVEWALDATAY-----SAKIETKSFPIGFVV---PGFNLITKV
1BMLc	SQLVVSV--LKFFEIDLT--LELLKAIQEQLIA--------DATITGKVYFATLPTQP--EFLLS-HV
1MG4	KKVRFYRNG-FGIVYAISPDFSFEALLADLTRTLNLP----TIYTILKKIS---SLDQLEGESYVCGS
1SE2	QNVLIRVN-KISFEVQTDKKVTLDIKARNFLINKNL---TGYIKFITFWY-----S---KSVKIEVHL
1AN8	HKLLGNLSG-QNLNIILEKDVTIDFKIRKYLMDNKI---SGRIEIGHEQI-----M---KNFHFDIYL
1AW7a	IELPLKVHG--KYWPKFDKKLALDFEIRHALTQIGL---GGYWKITTYQS-----I---DEITIEAEI
2IGD	TTYKLVIGK--ETTTKAV---DAEKAFKQYANDNGVD---GVWTYD---------------KTFTVTE
2PTL	VTIKANLANGQTAEFK-G---TATSEAYAYADTLDNG---YTVDVA----------------TLNIKF
1ACC	TTARIIF---VERRIAA---MTLKEALKIAF---GF-ITEFDFNFD----VLDIKL---AKMNILIRD
1G0Sa	HAAVLLP-----VAGMIEEGESVEDVARREAIEAGLI----RTKPV---------------SIMVGE-
1I9Aa	LAFSSWLF----VCGHPQLGESNEDAVIRRCRYEGV-----PPESI--------------CPVFAART

Table 5. Characteristics of the functional homologues included in MSA S1

Gene Id/smart	Organism	Accession no. in GenBank	Protein name	Residues range aligned
Raf-1h RBD*	*H. sapiens*	NP_002871	Raf-1	56-130
sptrembl|Q8IN00|Q8IN	*D. melanogaster*	NP_732773	Unknown/locomotion defect	360-430
ENSANGP00000001206/3	*A. gambiae*	XP_312542	RGS-like	362-432
ENSMUSP00000021945/2	*M. musculus*	NP_058038	RGS-14	303-374
ENSMUSP00000030984/3	*M. musculus*	NP_775578	RGS-12	962-1032
SINFRUP00000158915/2	*F. rubripes*	ND	Q8WX95/RGS-like	290-359
sptrembl|Q13878|Q138	*H. sapiens*	P15056	B-Raf/fragment	154-199
sptremblnew|BAB32131	*M. musculus*	AAH04757	A-Raf	19-79
sptrembl|Q98TC3|Q98T	*S. quinqueradiata*	BAB39747	Raf-1	54-129
swissprot|P09560|KRA	*X. laevi*	TVXLRF	Raf-1	56-130
sptrembl|Q8I086|Q8I0	*D. melanogaster*	AAN17541	Polehole	77-148
ENSANGP00000003633/1	*A. gambiae*	XP_318144	Raf-like	116-186
sptrembl|Q8MXT8|Q8MX	*C. elegans*	NP_741430	Lin-45/Raf-like	127-203
sptrembl|Q9GT28|Q9GT	*B. malayi*	AAG12472	Raf kinase/fragment	52-128
ENSANGP00000002581/1	*A. gambiae*	XP_316614	Tiam-1	1093-1160
ENSMUSP00000002588/7	*M. musculus*	NP_033410	Tiam-1	765-832
SINFRUP00000133952/7	*F. rubripes*	ND	Tiam-1	736-803
ENSMUSP00000024562/8	*M. musculus*	ND	NM_011878/tiam-2	831-903
ENSP00000275245/186-	*H. sapiens*	AAF05900	Tiam-2	186-257
sptrembl|Q8IN00|Q8#I	*D. melanogaster*	ND	CG5248-PC	431-501
ENSANGP00000001206#/	*A. gambiae*	XP_312542	RGS-like	433-504
ENSMUSP00000021945#/	*M. musculus*	NP_058038	RGS-14	376-446
ENSMUSP00000030984#/	*M. musculus*	NP_775578	RGS-12	1034-1104

*Raf RBD was not included in this alignment, because of high similarity with other sequences in the alignment. It is simply indicated to serve as a reference.

Table 6. Characteristics of the structural analogue included in MSA S2 and S3 (SCOP and SCOP-FSSP)

PDB code and chain aligned	Super-family* (SCOP)	Residues range correponding to ubi-like fold†	α-helix length‡	Ω-angle§	α-helix inner core residues¶	Volume of inner core residues sidechain‖	Amino acids at inner core Residues**
1RFA	**1**	**56-130**	**12**	**-50**	**I,I+4,I+8**	**601**	**I,V,L,L,L,V,L,V**
1UBI	1	1-71	12	-60	I,I+3,I+7	619	I,V,I,V,I,L,L,L
1A5R	1	22-92	12	-45	I,I+3,I+7	635	L,V,L,L,Y,F,I,V
1MG8	1	3-79	12	-60	I,I+3,I+7	601	L,I,V,L,V,L,V,L
1VCBa	1	3-79	12	-60	I,I+3,I+7	601	L,I,V,L,V,L,V,L
1M94a	1	2-72	12	-60	I,I+3,I+7	613	V,V,V,F,L,L,L,L
1J8Ca	1	33-102	12	-60	I,I+3,I+7	594	V,V,V,F,I,L,V,L
1H8Ca	1	4-80	12	-45	I,I+3,I+7	619	L,I,L,V,V,L,L,L
1JRUa	1	296-370	11	-45	I,I+3,I+7	646	I,I,I,I,I,L,Q,L
1EO6a	1	29-111	12	-45	I,I+3,I+7	599	V,V,V,F,I,L,V,Y
1EF1a	1	5-82	12	-60	I,I+3,I+7	542	V,V,G,L,V,L,F,F
1GG3a	1	1-77	13	-45	I,I+3,I+7	563	C,V,Q,L,C,I,F,F
1LFDa	1	17-99	12	-60	I,I+3,I+7	518	I,V,A,V,A,L,F,L
1E8Xa	1	219-309	12	-60	I,I+3,I+7	640	I,I,P,I,F,L,L,L
1K8Rb	1	71-147	12	-75	I,I+4,I+8	653	L,F,Y,L,L,V,L,I
1L7Ya	1	14-88	12	-45	I,I+3,I+7	604	F,I,F,V,A,I,L,L
1D4Ba	2	36-103	10	-45	I,I+4,I+8	575	F,V,L,A,L,L,L,V
1C9Fa	2	9-76	12	-15	I,I+4,I+8	567	V,L,L,G,F,L,L,L
1F2Ri	2	19-89	12	-60	I,I+4,I+8	511	C,L,L,A,L,L,F,A
1IP9a	2	13-84	12	-45	I,I+3,I+7	653	I,F,Y,L,I,L,I,V
1Q1Oa	2	762-854	15	-60	I,I+4,I+8	667	F,I,L,I,I,I,I,L
1FMAd	3	1-76	12	-60	I,I+3,I+7	557	I,V,V,L,M,A,V,F
1F0Za	3	1-61	8	-90	I,I+3,I+7	649	I,F,V,L,L,L,I,L
1JSBa	3	8-68	8	-60	I,I+3,I+7	594	F,V,I,V,L,V,I,V
1QF6a	4	2-61	8	-60	I,I+3,I+7	528	I,L,P,V,I,G,L,I
1JALa	4	280-361	8	-60	I,I+3,I+7	473	Y,T,A,A,I,A,M,F
1FRRa	5	2-90	8	-90	I,I+4	375	T,L,I,A,-,G,I,T
1I7Ha	5	2-102	8	-60	I,I+4	456	I,I,I,A,-,C,V,I
1AYFa	5	7-108	8	-90	I,I+4	531	V,F,L,V,-,L,V,V
1PUT	5	1-103	8	-60	I+4,I+8	449	V,Y,-,L,A,V,V,V
1L5Pa	5	1-93	8	-45	I,I+4	486	I,A,L,L,-,C,F,L
1E0Za	5	1-117	8	-60	I,I+4	458	V,Y,I,A,-,S,I,Y
1FEHa	5	2-76	8	-60	I,I+4	469	I,I,I,A,-,V,I,T
1HLRa	5	2-77	9	-60	I,I+4,I+8	619	K,I,L,L,L,V,I,T
1FO4a	5	6-89	9	-60	I,I+4,I+8	600	L,F,L,L,L,V,V,T
1JROa	5	1-79	8	-45	I,I+4	536	I,F,L,L,-,V,L,T
1FFVa	5	4-78	9	-60	I,I+4,I+8	571	I,V,L,L,L,V,V,T
2PIA	5	236-321	8	-45	I+4,I+8	536	F,V,-,I,L,T,L,L
1NENb	5	1-93	12	-90	I,I+3,I+7	612	L,F,V,A,L,M,I,I
1KF6b	5	4-92	12	-90	I,I+3,I+7	564	L,I,L,A,I,V,M,V

Table 6 (continued)

PDB code and chain aligned	Super-family* (SCOP)	Residues range correponding to ubi-like fold†	α-helix length‡	Ω-angle§	α-helix inner core residues¶	Volume of inner core residues sidechain‖	Amino acids at inner core Residues**
1QLAb	5	3-93	12	-90	I,I+3,I+7	638	L,I,I,V,I,I,I,L
1JQ4a	5	5-96	8	-90	I+4,I+8	388	I,A,-,V,A,A,I,L
1KRHa	5	4-97	8	-75	I,I+4	455	V,L,L,A,-,A,F,I
1C78a	6	23-131	13	-75	I,I+4,I+8	533	L,V,I,V,L,A,L,T
1BMLc	6	16-141	13	-75	I,I+4,I+8	500	L,V,L,I,L,A,L,-
1MG4	7	57-133	12	-60	I,I+3,I+7	621	V,F,F,L,L,I,Y,C
1SE2	8	129-234	16	-45	I,I+4,I+8	563	V,I,L,A,L,I,I,V
1AN8	8	100-206	16	-45	I,I+4,I+8	585	L,G,I,I,L,I,F,I
1AW7a	8	98-193	16	-45	I,I+4,I+8	639	L,L,L,I,L,W,I,A
2IGD	9	6-61	15	-30	I,I+4,I+8	592	Y,L,A,F,A,W,F,V
2PTL	9	18-76	16	-15	I,I+4,I+8	413	I,A,A,A,A,V,L,I
1ACC	other	487-593	9	-60	I,I+4,I+8	623	A,I,L,L,F,F,I,I
1G0Sa	other	58-148	13	-45	I,I+4,I+8	357	A,L,V,A,A,T,M,G
1I9Aa	other	33-125	13	-60	I,I+4,I+8	455	F,S,N,V,C,P,F,A

*SCOP database subdivides the β-grasp ubiquitin-like topology into 11 superfamilies: 1. ubiquitin-like 2. CAD/PB1 3. MoaD/ThiS 4. TGS-like 5. 2Fe-2S Ferredoxin-like 6. Staphylokinase/Strep-tokinase 7. Doublecortin (DC domain) 8. Superantigen toxins, C-terminal domain 9. Immunoglobulin-binding domains 10. Translation initiation factor IF3, N-terminal domain 11. glutamine synthetase N-terminal domain (this numbering is used in the table). The structures classified in "other" are categorized in other fold types, but display structure similar to the ubiquitin-roll topology (retrieved in the FSSP database). Note that a 12th superfamily was recently added to β-grasp ubiquitin-like, the TmoB-like. It is mostly similar to the 5 ubiquitin-like superfamilies and has only one member as it is now.

†Sequence range used in the alignment. Accounts for insertions or deletions inside the interval are mentioned.

‡The length of the α-helix was evaluated by inspection of the h-bond pattern of backbone atoms.

§Defined as the angle between the plane of the β-sheet (β-strand 5 is used as a guide) and the α-helix axis (11). The angle is approximated by manual observation of the structures.

¶The residues of the α-helix participating in the hydrophobic core are defined by visual examination of each structure. The numbering (e.g. I, I+3, I+4, I+7 and I+8) refers to positions 78, 81, 82, 85 and 86 of the alignments (for more information, see text under the subheading: Effect of the Variation in the Register of the α-Helix on the Entropy Scores).

||Volume of inner core residues side chains in cubic angstrom (Å^3) were calculated according to the volume of amino acid chemical groups as defined by Richards *et al.* (12). The core residues were defined by visual evaluation of the structure and the entropy profile. They correspond to residues 58, 60, 78, 81 or 82, and 85 or 86, 98, 126 and 128 of the Raf RBD. Although there is diversity of the sequences and variations in fine details of the structure, the packing of the α-helix involved these residues most of the time. Nevertheless, it is observed quite frequently that residues 78 and 86 (core position 1 and 3 of the α-helix) are not true core residues.

**List of amino acid observed at inner core residue as described above in the two last footnotes and by taking into account the difference in α-helix packing. Note that structures with shorter α-helix have gaps at core position (78 or 86).

Table 7. Entropy profile characteristics and mean pairwise sequence identity for the experimental data and the 3 natural sequence alignments

Alignments	No. of sequences, *N*	Mean pairwise identity*,%	Highest pairwise identity†,%	Mean entropy‡	Standard deviation‡
Experimental	Table 1	ND	ND	0.82	0.14
SMART (MSA S1)	22	20.6	78.0	0.55	0.16
SCOP (MSA S2)	27	10.3	35.4	0.71	0.10
SCOP-FSSP (MSA S3)	54	9.4	35.4	0.77	0.09

*The pairwise amino acid identity is calculated for all pairs of sequences from the natural sequence alignments (Table 4). Positions displaying gaps in either of the two sequences are not considered in the calculation to avoid biasing the result.

†The highest pairwise identity observed within an alignment.

‡The entropy was calculated according to Eq. **1**. The mean entropy and its standard deviation were calculated by considering each position degenerated in the main experiment or each position in the natural sequence alignments plotted (Fig. 2 *B-D* and *Methods*).

Table 8. Raw z scores of occurrences for the 20 amino acids at all residues degenerated experimentally

	S	N	T	I	R	V	F	L	P	N	K	Q*	R	T	V	V	N	V	R	N	G	M	S
W	-0.7	1.4	0.7	-1.4	-0.7	-1.4	1.4	-1.4	-1.4	-0.7	-0.7		0.5	-1.4	-0.8	-1.4	-0.8	-1.4	-0.3	0.1	-1.0	-1.0	-1.4
Y	-0.7	-0.7	0.7	-1.4	0.0	-1.4	0.7	-1.4	-1.4	-0.7	-0.7		1.2	-1.4	0.5	-1.4	-0.1	-1.4	-0.6	-1.0	-0.3	-0.3	-1.0
F	0.7	0.0	-0.7	0.7	1.4	4.1	0.7	0.8	-1.4	0.1	0.1		1.8	-1.4	-0.8	-0.8	-0.1	-0.3	-0.3	-1.4	-1.4	-1.4	-1.4
V	-1.0	-1.0	0.5	2.0	0.0	10.9	-1.0	-1.5	-2.0	-1.5	-1.0		1.2	-1.6	1.2	3.1	0.7	7.3	-1.2	-1.2	-1.2	0.7	-0.9
I	-1.4	-1.4	-0.7	4.8	0.7	4.8	-0.7	6.7	-0.7	-1.4	-1.4		1.2	-0.1	3.1	1.2	1.2	3.2	-0.6	-0.3	-1.4	1.6	-1.4
L	0.4	-2.1	1.2	11.5	-0.9	-2.5	-0.9	15.1	-1.2	-1.2	-1.6		0.6	-2.5	-0.2	0.9	0.2	-1.1	0.7	-0.2	-1.1	0.2	-1.4
M	0.7	4.1	1.4	2.1	0.0	-1.4	1.4	1.5	-1.4	-0.7	0.1		1.2	-0.8	2.4	2.4	1.8	-1.0	0.1	-0.6	-1.0	2.0	-0.3
C	0.0	2.1	2.1	0.0	0.7	2.7	-1.4	0.1	-0.7	-1.4	0.8		-0.1	-0.1	0.5	-0.1	-0.1	1.3	0.1	0.5	-1.0	0.1	0.1
A	1.5	0.0	0.5	-2.0	2.5	4.4	2.5	-2.0	-1.0	-1.0	0.6		0.3	-0.2	-0.6	-0.2	-0.6	3.7	-0.4	1.0	1.5	0.2	-0.9
G	0.5	-1.0	-1.5	-2.0	-2.0	-2.0	-1.5	-2.0	4.9	-1.0	1.7		-1.6	-1.6	-2.0	-1.6	-1.1	-1.2	0.2	-0.1	2.4	-0.1	-1.2
P	0.0	0.0	0.5	-2.0	0.0	-2.0	-2.0	-2.0	16.0	-2.0	-2.0		-2.0	-2.0	-1.6	-1.6	-1.1	1.5	1.0	3.7	1.8	-1.2	0.4
T	1.5	1.0	1.5	-1.0	2.5	-1.5	-2.0	-2.0	-1.0	-1.0	0.6		-2.0	12.7	4.0	1.2	0.3	-0.9	0.4	-0.1	0.2	-0.4	5.6
S	-0.9	2.0	-1.3	-2.1	-0.9	-2.5	0.8	-2.5	-2.1	1.0	1.4		-1.0	5.9	-1.0	-1.0	0.6	-2.1	0.2	0.4	1.1	0.4	3.6
N	0.0	0.0	0.7	-1.4	0.7	-0.7	1.4	-1.4	0.1	5.2	3.0		-0.1	-1.4	-0.1	-0.8	-0.8	-1.0	0.1	-0.6	0.9	-0.6	-0.3
Q	-0.7	0.0	0.0	-1.4	-1.4	-1.4	0.7	-1.4	0.1	-0.7	0.1		0.5	-0.8	3.1	2.4	1.8	-1.0	0.1	-1.0	-0.3	0.9	-1.0
D	0.0	0.0	-1.4	-0.7	-1.4	-0.7	0.7	-1.4	-1.4	7.4	0.8		-0.8	-1.4	-1.4	-1.4	-0.8	-1.0	-0.3	-1.4	-1.0	-1.0	-1.0
E	-1.4	0.0	-1.4	-1.4	1.4	-1.4	-0.7	-1.4	-1.4	5.2	1.5		-1.4	-1.4	-0.1	-1.4	-1.4	-0.3	-1.0	0.5	0.9	0.9	-0.6
K	0.0	-1.4	0.0	-1.4	1.4	-1.4	2.1	-1.4	-1.4	-0.7	-1.4		1.2	-0.8	-1.4	1.2	-0.8	-0.6	1.3	0.9	0.5	0.1	2.0
R	1.6	0.0	-0.5	-2.5	-0.5	-2.5	-0.5	-2.5	-2.5	-1.6	-1.2		1.3	-2.5	-1.0	0.9	1.3	-2.1	0.0	0.7	0.4	0.4	0.2
H	0.0	-0.7	-0.7	-1.4	-1.4	-1.4	2.1	-1.4	-1.4	1.5	1.5		0.5	0.5	-1.4	-1.4	0.5	-0.6	1.3	-0.3	-0.6	-0.6	-0.6

Table 8 (continued).

	L	H	D	C	L	M	K*	A	L	K	V	R*	G	L	Q	P	E	C	C	A	V	F	R
W	-1.4	-1.4	-0.8	-1.4	-1.4	-1.4		-1.4	-1.4	0.8	-0.3		-0.7	-0.1	-1.4	-0.1	-1.4	-1.4	-1.4	-0.9	-1.4	0.1	1.1
Y	-1.4	1.7	-0.8	-1.4	-1.4	-0.8		-1.4	-1.4	-0.3	-0.3		-1.4	1.3	-1.4	-0.7	-1.4	1.3	-1.4	0.6	-0.4	1.1	1.1
F	-1.4	-0.8	-1.4	-0.8	-1.4	-0.8		-1.4	3.0	0.8	-0.3		-1.4	-0.1	-0.7	-0.1	-1.4	-0.7	-1.4	-0.9	-1.4	0.1	1.1
V	5.2	-0.2	-0.2	-0.2	1.1	-0.2		-0.4	-0.8	2.4	0.8		-1.5	0.9	-0.1	1.8	-0.6	-1.5	-1.7	0.5	9.0	0.1	0.1
I	3.0	-0.2	1.1	6.7	2.3	-0.2		-1.4	4.1	-0.8	0.3		-1.4	3.3	-0.1	-0.1	-1.4	-1.4	0.1	-0.4	6.5	1.6	1.1
L	15.4	-1.4	-2.5	0.8	13.9	-0.3		-2.5	10.1	-0.2	1.1		-2.1	8.3	-1.3	0.3	-2.1	-2.1	7.5	-0.5	0.7	0.4	-0.5
M	-1.4	-0.2	0.5	1.1	6.1	1.1		-1.4	8.0	0.8	1.9		-0.7	3.3	0.6	-0.7	-0.7	-1.4	10.4	1.1	-0.4	-0.9	0.6
C	-1.4	-0.8	-1.4	3.6	-0.2	-0.2		0.3	0.3	0.3	-0.3		-1.4	1.3	-0.1	-0.7	-0.7	0.6	-0.9	0.1	-0.4	-0.9	-0.9
A	-1.6	0.2	0.2	-0.2	-2.0	0.2		2.4	-1.2	0.0	-2.0		-0.1	-1.5	-0.1	1.4	0.4	0.9	-1.3	-0.2	-0.2	0.1	-0.6
G	-2.0	-1.1	-0.7	-1.6	-2.0	-0.2		-0.8	-2.0	0.8	-1.6		7.6	-1.5	-0.1	-0.1	1.4	-0.6	-1.3	0.1	-1.0	0.1	-1.3
P	-2.0	-2.0	-2.0	-2.0	-2.0	-2.0		1.2	-2.0	-2.0	-2.0		-2.0	-2.0	-0.1	0.4	-0.6	-2.0	-2.0	-2.0	-1.0	0.5	-1.7
T	-2.0	-2.0	1.6	2.0	-2.0	-1.1		0.0	-2.0	-1.6	-0.4		-1.1	-1.1	-0.1	-1.1	-0.6	-0.1	2.9	1.9	-0.6	1.2	-1.0
S	-2.5	0.1	-1.8	0.5	-2.1	0.5		4.8	-2.5	-1.2	0.8		0.7	-2.1	0.7	0.3	-0.1	0.3	-1.9	1.0	-2.5	-0.7	-0.2
N	-1.4	-0.2	2.3	1.7	-1.4	-0.2		-1.4	-1.4	0.8	1.9		4.0	-1.4	-0.1	-1.4	3.3	-0.1	0.1	-1.4	-1.4	0.6	-0.9
Q	-1.4	2.3	1.1	-0.8	-1.4	1.7		-0.8	-1.4	-0.3	0.3		0.6	-1.4	1.9	0.6	3.3	2.6	-1.4	-1.4	-0.9	0.1	-0.4
D	-1.4	-1.4	6.1	-1.4	-1.4	-0.2		-1.4	-1.4	-0.8	-1.4		1.9	-1.4	4.6	-1.4	1.3	7.3	-0.9	0.6	-0.4	-1.4	-1.4
E	-1.4	4.2	3.6	-0.2	-1.4	3.6		-1.4	-1.4	0.8	-0.8		-0.1	-0.7	-0.1	-0.7	4.6	3.3	-1.4	1.6	-0.4	-0.9	-0.4
K	-1.4	3.0	1.7	-1.4	-1.4	1.1		5.2	-1.4	2.5	0.8		1.9	-1.4	0.6	-0.1	1.9	-0.7	-0.4	2.5	0.1	-0.4	-0.4
R	-2.5	3.1	-0.7	-2.5	-2.5	0.1		0.8	-2.5	0.1	0.8		-0.5	-1.7	-0.1	1.1	-1.3	-0.9	-1.9	0.1	-1.3	0.1	5.4
H	-1.4	-0.8	-0.8	0.5	-1.4	1.7		-0.3	-1.4	-0.3	2.5		-0.7	-1.4	-0.7	0.6	-0.1	1.3	-1.4	-1.4	-1.4	0.1	-0.9

Table 8 (continued).

	L	L	H	E	H	K	G	K	K	A	R	L	D	W*	N	T	D	A	A	S	L	I	G
W	0.6	-0.8	-0.8	0.4	-1.4	-0.8	-1.4	-1.4	0.8	-0.9	-0.9	-1.4	-1.4		-1.4	-0.8	-0.8	-0.8	-0.8	-1.4	-1.4	-1.4	-1.4
Y	-0.9	-0.8	1.0	0.4	-0.2	-0.8	-1.4	-0.2	-1.4	-0.3	-0.9	-0.9	0.2		-0.8	-0.8	-1.4	-1.4	-0.2	-1.4	-0.1	-1.4	-0.8
F	-0.9	-1.4	-0.2	-1.4	-0.8	-0.2	0.4	-0.2	-1.4	-0.9	-1.4	-0.9	-1.4		-0.2	-0.8	-0.8	-1.4	-0.8	-0.2	-1.4	-1.4	-1.4
V	0.8	-2.0	-0.3	-1.1	0.2	-1.1	0.6	0.6	-0.4	1.9	-1.2	4.7	-0.8		-2.0	0.7	-1.6	0.2	-1.1	-0.7	0.8	0.3	-0.1
I	0.6	0.4	0.4	-0.2	0.4	0.4	0.4	-0.8	1.3	1.3	-0.9	1.3	-1.4		-1.4	-0.2	-0.8	-0.8	-0.8	-1.4	2.5	-0.8	-1.4
L	1.0	-1.4	0.0	-1.1	-0.3	-0.3	-1.1	-0.3	0.1	0.4	-0.6	7.0	-1.2		-2.5	-1.0	-2.5	-2.5	-2.5	-2.1	6.0	0.6	-1.7
M	1.6	1.0	-0.8	-0.8	-0.8	-0.2	1.0	-0.8	1.3	1.3	0.8	0.2	-0.9		-0.8	4.2	-0.8	-0.2	1.1	-0.8	1.9	0.5	-1.4
C	-0.4	-1.4	0.4	-1.4	2.2	0.4	-0.2	-0.8	-0.9	-0.3	-0.3	-1.4	-0.9		-0.2	3.0	0.5	4.2	-0.8	-0.8	4.5	0.5	0.5
A	-0.2	0.6	-0.3	-0.3	-1.1	-0.3	-0.7	2.8	-1.6	2.3	1.5	1.1	1.1		0.2	0.2	-0.7	10.1	-0.7	0.7	-1.5	-0.1	-1.5
G	-0.6	1.0	-0.7	2.8	-0.7	1.9	-0.3	0.2	0.0	-0.4	-0.8	-2.0	0.7		2.5	-0.7	-1.1	0.7	-0.7	0.2	-1.5	1.7	8.3
P	-1.7	0.6	-0.3	-0.7	0.6	-0.7	0.2	-0.7	-1.2	-0.4	-0.4	-1.2	0.7		-1.1	-2.0	3.4	-0.7	-1.6	-1.1	-2.0	-0.6	-0.1
T	-0.6	0.2	-0.7	1.5	-1.1	-1.1	1.0	0.2	1.1	-0.4	1.1	0.7	0.3		-0.2	1.6	1.1	-0.2	2.9	1.6	-1.5	1.3	0.3
S	-1.3	0.4	0.4	-1.1	0.4	0.0	1.5	0.4	0.1	-1.5	3.4	-1.9	1.7		1.2	0.5	-0.3	0.5	3.8	1.6	-1.3	-0.6	-0.6
N	-0.4	1.6	1.6	1.6	0.4	-0.2	1.0	0.4	0.8	0.2	-0.3	-1.4	-0.3		1.7	-0.8	2.3	-0.2	-0.2	-0.8	-0.1	-0.1	1.2
Q	-0.4	1.0	-0.2	0.4	0.4	-0.8	-1.4	1.0	-0.9	-0.3	-0.3	-0.3	3.0		1.7	1.7	0.5	-1.4	-0.2	-0.2	1.2	1.9	1.9
D	-0.9	-0.2	2.2	0.4	-0.2	0.4	1.0	1.0	-0.9	0.2	-0.9	-0.9	2.4		3.0	-0.2	9.2	-0.8	1.1	5.5	-1.4	0.5	3.2
E	0.6	4.7	2.2	1.0	1.6	2.8	1.6	1.0	0.2	1.9	2.4	-0.9	0.8		0.5	-0.8	1.7	-1.4	4.2	2.3	-0.8	-1.4	-0.8
K	2.1	-0.8	-0.2	1.6	1.6	1.6	0.4	1.6	0.2	-1.4	0.2	-0.9	-0.3		1.7	-0.2	-0.2	-1.4	-0.2	1.1	-0.8	-1.4	-1.4
R	2.8	-0.7	-0.7	0.4	0.7	0.4	-0.3	-1.4	2.1	-0.6	-0.6	-2.2	-1.2		-0.7	-1.0	-2.1	-1.4	-0.3	-0.7	-1.7	2.1	-1.0
H	-0.9	0.4	-0.8	-0.8	-0.2	0.4	-1.4	-1.4	1.3	-0.9	-0.3	-0.9	-0.3		1.7	-0.2	-1.4	-1.4	-1.4	0.5	0.5	-1.4	-0.8

Table 8 (continued).

	E	E	L	Q	V	D	F
W	-0.1	-1.4	-1.4	0.0	-0.7	-0.7	-1.4
Y	-0.8	-0.7	-1.4	1.4	-0.7	-0.7	-0.7
F	0.5	-1.4	-0.7	1.4	-1.4	-1.4	-0.7
V	-2.0	-2.0	1.5	2.0	16.1	0.5	1.0
I	-0.8	-0.7	7.7	0.0	5.6	0.0	0.7
L	-0.6	-1.3	10.5	-1.3	0.0	-0.8	-0.4
M	-1.4	0.0	0.0	0.7	0.0	0.0	0.7
C	1.2	0.0	0.0	0.7	0.0	0.0	-0.7
A	0.8	-1.0	-2.0	-0.5	-2.0	-2.0	0.5
G	-0.1	-1.5	-1.5	-2.0	-1.5	-2.0	-2.0
P	-2.0	-0.5	-2.0	-2.0	-1.5	-2.0	0.0
T	4.1	1.5	-2.0	1.5	-1.5	2.5	-0.5
S	1.4	1.7	-1.3	-0.4	-2.1	-0.8	1.7
N	1.2	0.7	-0.7	0.7	-0.7	0.0	1.4
Q	-0.1	-0.7	-1.4	0.0	-1.4	1.4	2.1
D	-0.8	2.8	-1.4	0.7	-1.4	3.5	0.7
E	1.9	0.0	-0.7	2.1	-1.4	6.3	2.8
K	-0.8	3.5	-0.7	0.7	-1.4	0.7	-0.7
R	-1.7	1.3	-2.1	-1.7	-2.5	-0.4	-2.1
H	1.9	1.4	-1.4	0.0	-0.7	0.0	0.0

*Grey columns indicate unvaried residues

Table 9. Listing and classification of significant amino acid selections distribution ($p < 0.05$) observed experimentally and in the three natural sequence alignments

Residue*	Experimental†	SMART (MSA S1)†	SCOP (MSA S2)†	SCOP-FSSP (MSA S3)†	Role‡
56	M, C, S	K, S	M, I, C	M, I, H	G/oC
58	L, **I**, M	C, F	I, V, F, C	I, V, L, F	G/C
60	**V**, I, A, F, C	V, L	V, I, F	V, I, F	G/C
62	**L**, I	L	T	V, T, L	G/oC
63	**P**, G	P	N, P	N, P	S/t
64	D, **N**, E	N, D	-	D	S/t
66	ND (Table 3)	Q	-	Q	S/B
68	**T**, S	T, V, C	F	F, Y	S/B
70	**V**, M, Q	V	V, L	V, F, L, I	G/oC
72	**V**, A, I	V	V, C, A	V, A, C	G/oC
75	**G**	G	D, S	G, E	G/t
76	**M**	M, K, E, H	D, T	T, M	G/-
77	**T**, S, K	T, R, S	T, S	T, S	G/Nh
78	**L**, V, I	V, L, I	L, V	L, I	G/C
80	**D**, E, N	D, E	D, Q	D, K	G/Nh
81	I, **C**, T	V, C, A	L, V	A, V	G/oC
82§	**L**, M, I	L	L/**L,V**	K, L/**L, A, V, I**	G/C
83	E	E	E, C, A	R, E	G/-
84	ND (Table 3)	K, P	V	Q	S/B
85	K, S, **A**	I, A, L	I, L	A, I	S, G/Bo, C
86§	**L**, M, I, F	C, L	-/**I, L**	L, A/**L, I**	G/C
89	ND (Table 3)	R, Y	-	-	S/B
90	**G**, N	G, N, Q	G	G, N	G/Ch
91	**L**, M, I	L, I	L	L, I	G/Ch, oC
96	M, L, T	C, Y, H, V	F, L	C	G/oC
98	**V**, I	V, L	L	L, V, I	G/C
100	**R**	L, R	V, F	V, F, I	S/oC, mC, Sb

Table 9 (continued).

Residue*	Experimental†	SMART‡	SCOP‡	SCOP-FSSP‡	Role§
112	**L**, V	L, M, I	L, V, I	V, L	G/oC, mC
113	Q, **D**	D, C	D	D, L	G/t
114	ND	W, L	D, W	D, A	S/C
117	**D**, P, N	P, D	T, P	P, T	G/t
118	**A**, C	S, C, V	L, V	L, M, V, I	G/oC
121	**L**, C, I	L	C, L	C	G/oC, mC
123	**G**, D	G	G, D	G, D	G/t
125	K, D	E, D	T	V, T	S/mC,Sb
126	**L**, I	L, V	L, F, I, V	I, F, L, V	G/C
128	**V**, I	V, I, Y, L	L, F, V	L, V, I, T	G/C
129	E, **D**, T	E, D, H	Y	H	S/Sb

*Residues with amino acid selections distribution with $p < 0.05$ for at least one amino acid are listed in this table if the selection is also observed for the same amino acid type or class in either the structural analogue alignments or the functional homologue alignment.

†The amino acid type frequencies with $p < 0.05$ are presented from the most to the least significant deviation. In the experimental data set, the amino acids in bold are observed in the wild-type Raf RBD.

‡The residue with significant amino acid selections are classified on the basis of the convergence between the experimental data and either the functional homologues (S) or the structural analogues (G). By observation of the structure and consultation of the literature (10, 17, 18) residues are further described with one of the following acronyms: inner core (C), outer core (oC), mini core in carboxyl terminus (mC), turn (t), α-helix N-cap (Nh), α-helix C-cap (Ch), putative salt bridge (Sb), binding to *h-ras* (B). To simplify the descriptions in the article the outer core also included three residues classed in the mini core in this table: R100, L112, and L121. The residues that participate in the mini core and the organization of their side chains vary across structures included in the structural analogue alignments. Several structures do not even display this mini core (e.g. 2IGD and 2PTL).

§Two sets of amino acid (set1/**set2**) for which significant amino acid selections were observed are presented in this table and are based on: (*i*) The z score for position 82 and 86 of the SCOP and SCOP-FSSP alignments as displayed in Fig. 3 and (*ii*) a composite z score calculated by aligning inner core residues 2 and 3 of the α-helix, thus correcting for the two types of α-helix inner core residues register observed across the ubiquitin superfold (see text under the subheading: Effect of the Variation in the Register of the α-Helix on the Entropy Scores).

Table 10. Statistics concerning solvent accessibility and B factors distribution for inner versus outer core and all other residues

Structural characteristics and structure coordinates utilized	Types of residues	Average	Standard deviation
Solvent accessibility based on 1RFA	All other	78.8	44.7
	Inner	2.3	2.1
	Outer	19.2	19.7
Solvent accessibility based on 1GUA	All other	75.2	47.8
	Inner	1	1.3
	Outer	15.9	24.8
B factors for main chain atoms based on 1GUA	All other	17.2	8.7
	Inner	11.6	1.2
	Outer	13.3	3.0
B factors for lateral chain atoms based on 1GUA	All other	21.0	10.7
	Inner	12.1	2.3
	Outer	15.3	5.8

Table 11. Binding affinity of Raf RBD variants for *h-ras*

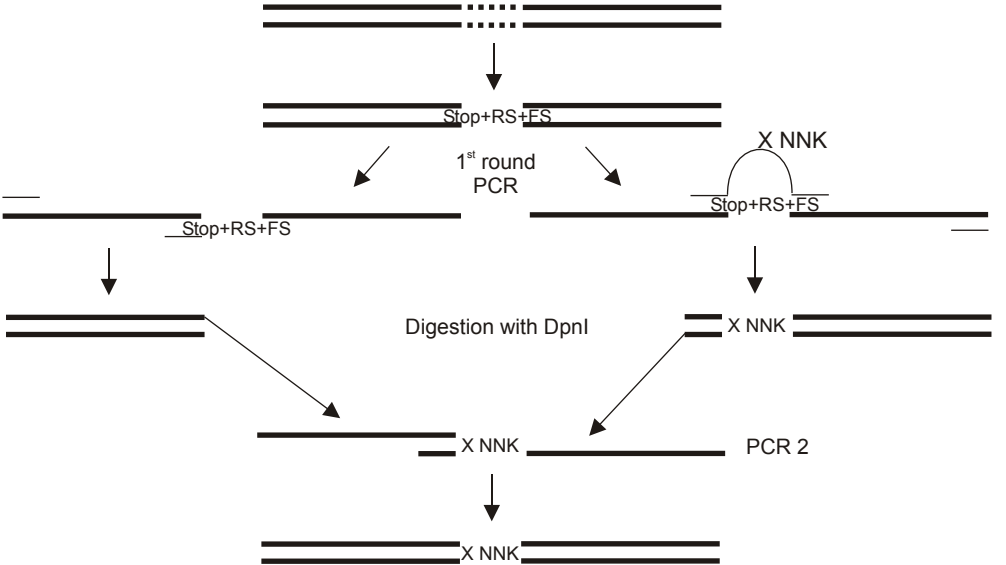
Libraries	Clone names	Sequences*	K_d†‡§, μM
wt	-	-	0.11 ± 0.029
S1	CF3	SFNLAVK	4.6
S2	A2	LPRLQ	3.6
S2	B2	LPGHQ	0.12
S2	H2	FTDGQ	8.6
S3	A3	GTRVT	7.4
S3	F3	WSIQR	0.87
S6	D6	QPLRLR	4.3
S6	E6	KALQQR	16
S6	G6	KRMTAR	5.4
S7	G3	EINHLQ	4.2
S8	H7	KLVIAR	0.42
S8	F8	CKLMRR	3.0
S9	A9	PDQSATG	0.72
S9	D9	KTQGCNG	0.11
S9	E9	TSGRVLH	0.07
S10	G9	VAILDW	3.7
C96M	-	-	0.38
R100A	-	-	0.64
D117A	-	-	0.13
E125A	-	-	0.76
D129A	-	-	0.74

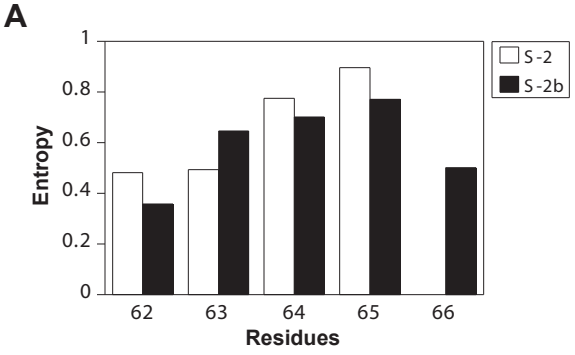
*Sequences of the clones in the degenerated region.
†According to *Methods*.
‡The dissociation constant for the wt Raf RBD was determined independently from four experiments (each in triplicate). The standard deviation is indicated.
§The clone D6 was isolated from library S6 after one round of selection. An unexpected point mutation occurred: K84Q.

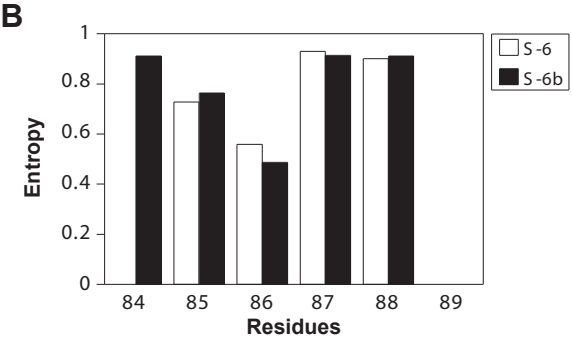
Table 12. Kinetic and thermodynamic parameters for characterized mutants of the Raf RBD

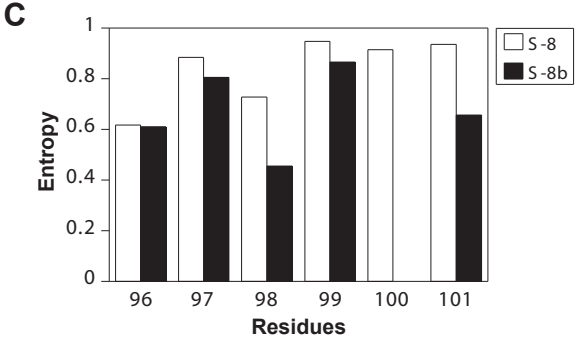
Libraries and clones	Sequence	k_f, s^{-1}	k_u, s^{-1}	m_f, kCal	m_u, kCal	ΔG, kCal. M^{-1}	M, kCal	C_m, M	β_t
wt	--	347	0.19	0.67	0.28	-4.44	2.37	1.87	0.70
V60A	--	66.0	0.43	1.05	0.26	-2.99	3.24	0.92	0.80
S1 (A1)	PWVDLDA	1220	4.01	0.61	0.23	-3.39	2.08	1.62	0.73
S1(CF3)	SFNLAVK	321	1.22	0.77	0.29	-3.30	2.62	1.26	0.73
S2 (A2)	LPRLQ*	179	0.40	0.86	0.27	-3.62	2.82	1.28	0.76
S2 (B2)	LPGHQ*	185	0.12	0.68	0.29	-4.34	2.41	1.80	0.70
S2 (H2)	FTDGQ*	1405	0.52	0.60	0.28	-4.69	2.18	2.15	0.68
S3 (A3)	GTRVT	242	0.39	0.71	0.30	-3.82	2.50	1.52	0.70
S6 (A6)	K*KLSER*	128	0.18	0.93	0.32	-3.87	3.09	1.25	0.74
S6 (C6)	K*RLVWR*	361	1.49	0.92	0.28	-3.25	2.96	1.10	0.77
S6 (G6)	K*RMTAR*	24.3	0.17	0.91	0.30	-2.93	2.99	0.98	0.75
S7 (B7)	SMSSES	183	0.72	0.93	0.25	-3.28	2.91	1.13	0.79
S7 (G7)	EILPGQ	379	0.77	0.93	0.25	-3.67	2.91	1.26	0.79
S7 (G3)	EINHLQ	249	1.85	0.79	0.29	-2.91	2.69	1.08	0.73
S8 (A8)	DKLGIW	230	0.88	0.84	0.27	-3.30	2.75	1.20	0.76
S8 (F8)	CKLMRR	453	11.4	0.88	0.36	-2.18	3.80	0.71	0.71
S10 (C10)	KRTVSW*	386	1.57	0.74	0.27	-3.26	2.50	1.31	0.73
S12 (H11)	LADC	541	1.82	0.79	0.33	-3.37	2.77	1.22	0.71
S13 (C12)	TSCHDL	74.1	0.76	0.84	0.25	-2.71	2.70	1.00	0.77
S13 (D12)	RLMVSD	95.3	0.81	0.93	0.24	-2.82	2.91	0.97	0.79
S13 (H12)	QLLLEF	232	0.16	0.60	0.28	-4.30	2.18	1.97	0.68

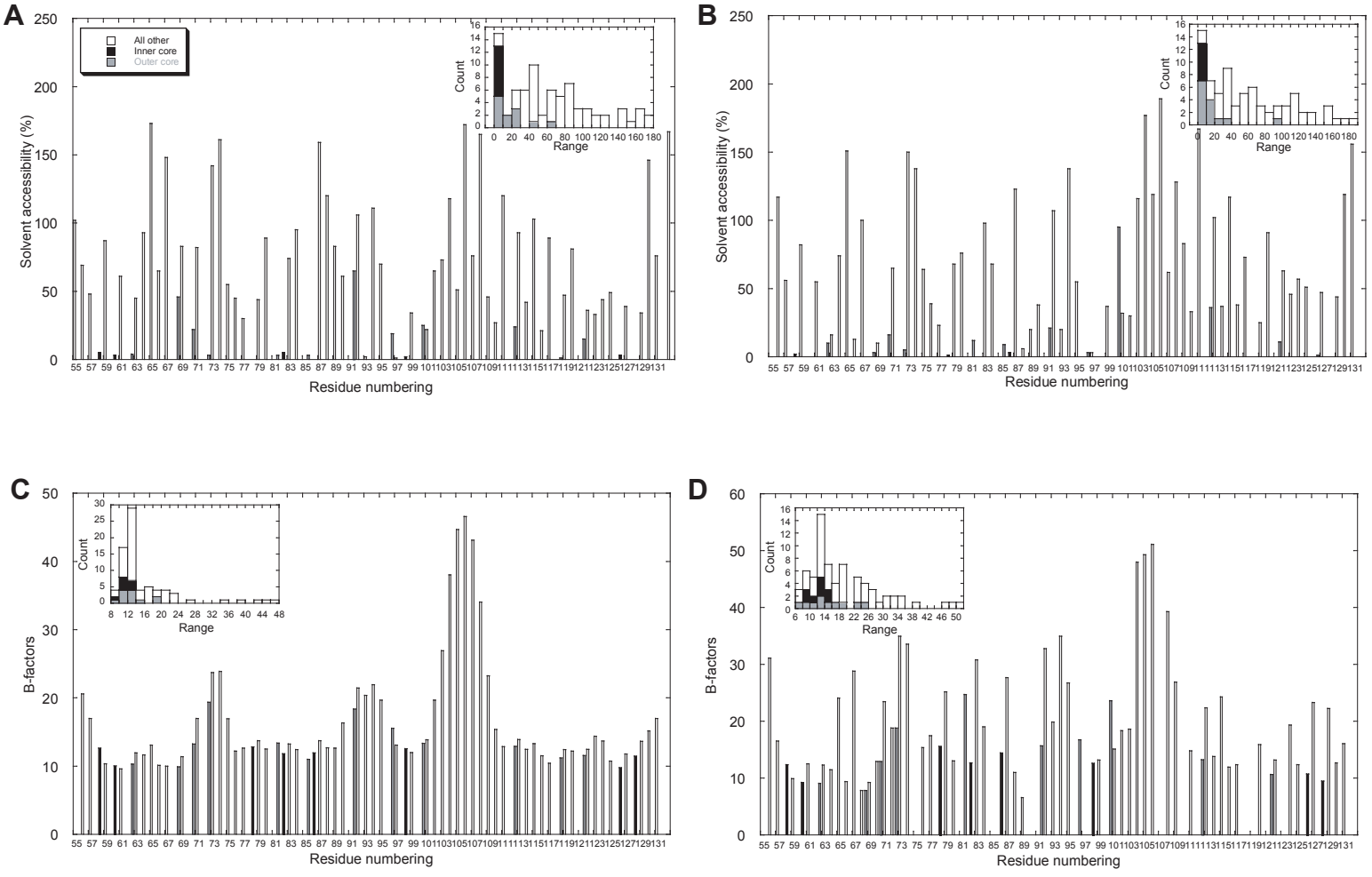
*Residues marked with an asterisk were not varied in the experiment.


Stop+RS+FS
1st round
PCR
X NNK
Stop+RS+FS
Stop+RS+FS
Digestion with DpnI
X NNK
X NNK
PCR 2
X NNK

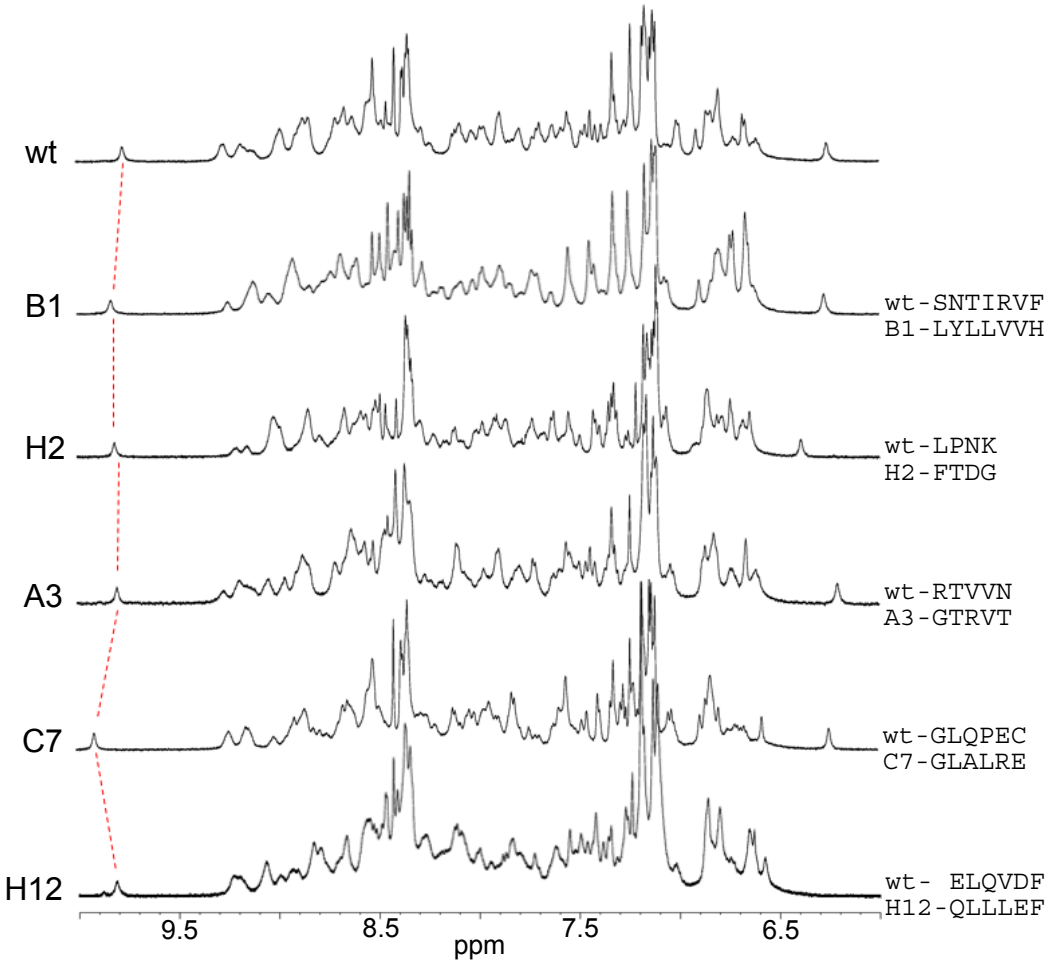

A
1
0.8
0.6
0.4
0.2
0
Entropy
S-2
S-2b
62
63
64
65
66
Residues


B
1
0.8
0.6
0.4
0.2
0
Entropy
S-6
S-6b
84
85
86
87
88
89
Residues

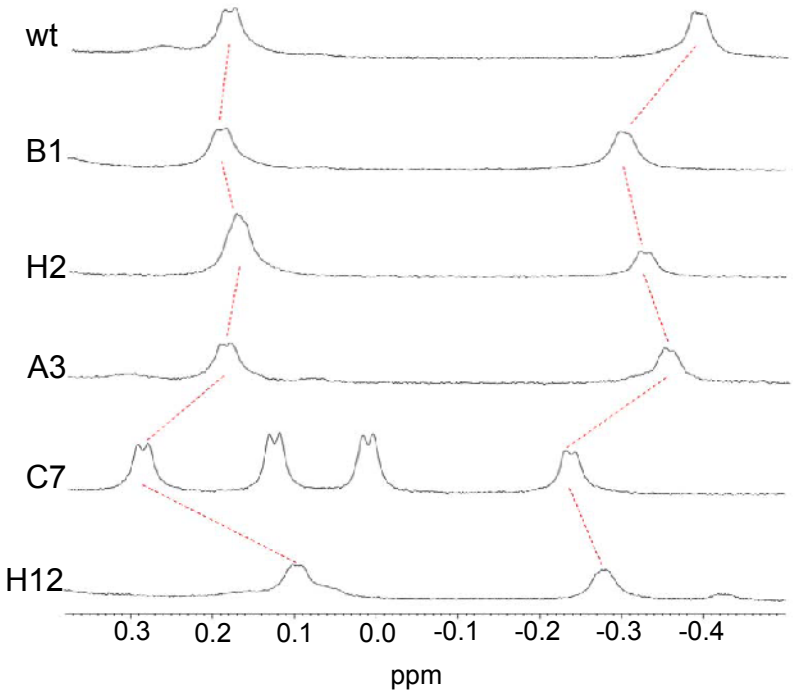

C
1
0.8
0.6
0.4
0.2
0
Entropy
S-8
S-8b
96
97
98
99
100
101
Residues


A
All other
Inner core
Outer core
Solvent accessibility (%)
Residue numbering
Count
Range
B
Solvent accessibility (%)
Residue numbering
Count
Range
C
B-factors
Residue numbering
Count
Range
D
B-factors
Residue numbering
Count
Range

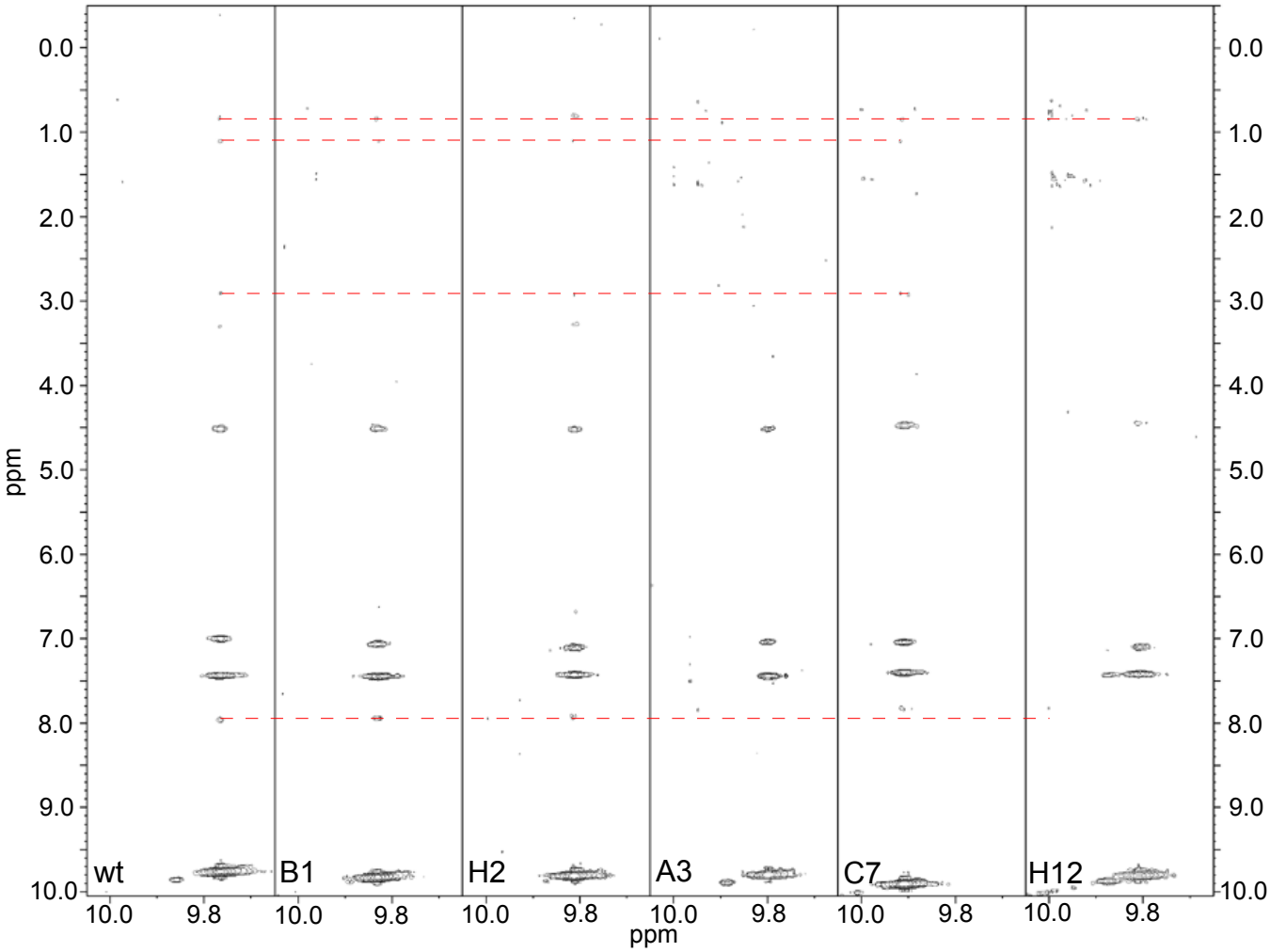
A


wt
B1
H2
A3
C7
H12
wt-SNTIRVF
B1-LYLLVVH
wt-LPNK
H2-FTDG
wt-RTVVN
A3-GTRVT
wt-GLQPEC
C7-GLALRE
wt- ELQVDF
H12-QLLLEF
9.5
8.5
7.5
6.5
ppm

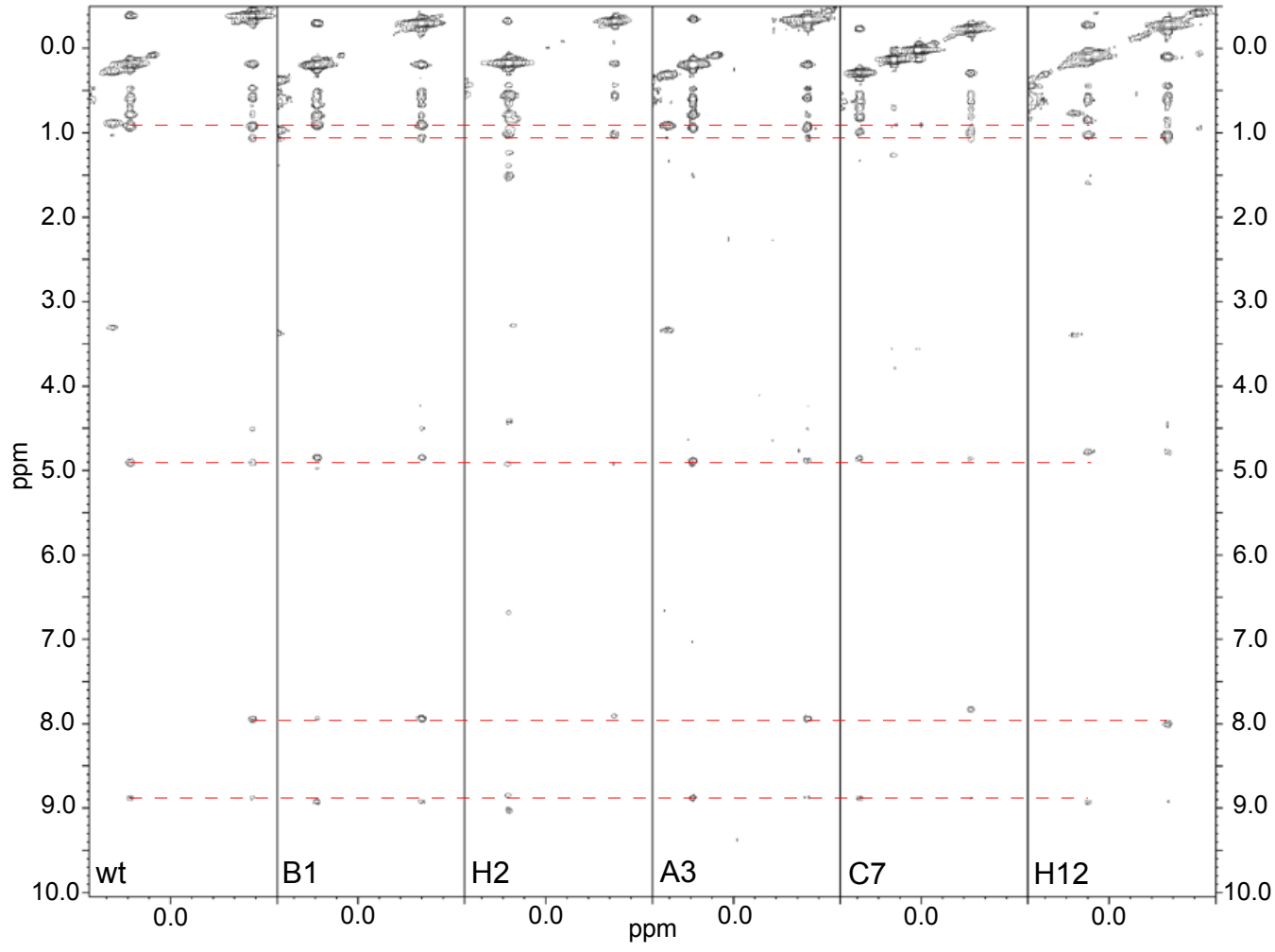
B

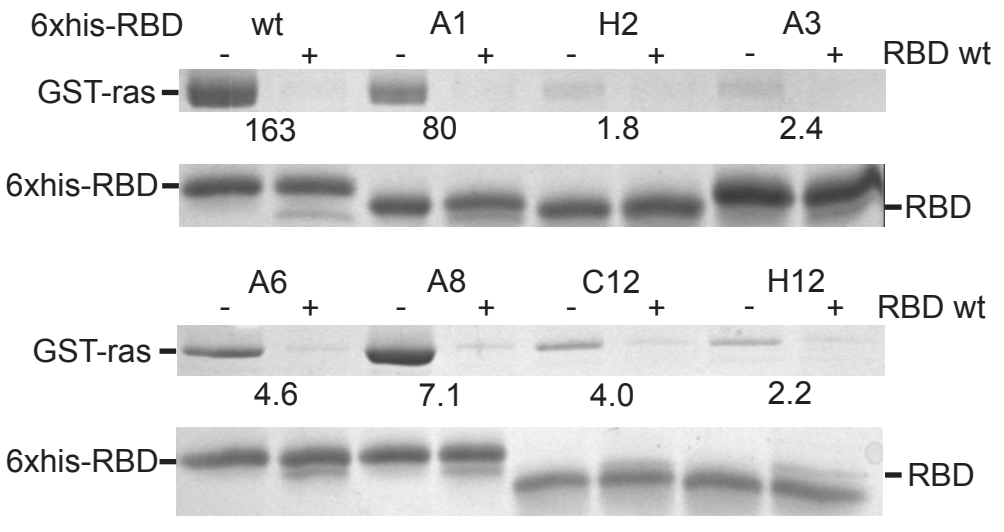

wt
B1
H2
A3
C7
H12
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
ppm

C



D




6xhis-RBD
wt
A1
H2
A3
- + - + - + - +
RBD wt
GST-ras
163
80
1.8
2.4
6xhis-RBD
RBD
A6
A8
C12
H12
- + - + - + - +
RBD wt
GST-ras
4.6
7.1
4.0
2.2
6xhis-RBD
RBD