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.

www.pnas.org/cgi/doi/10.1073/pnas.0500465102

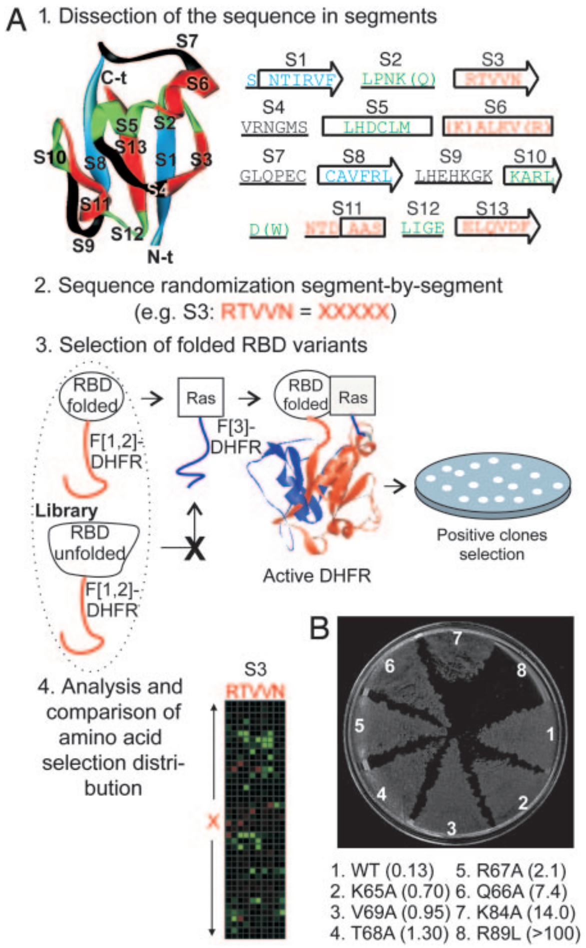

Fig. 1. Description of the sequence perturbation methodology. (*A*) Experimental strategy. (*1*) The Raf RBD is subdivided into 13 segments based on topological elements. Residues in parentheses were unvaried in the experiments. (*2*) Each segment is degenerated separately by PCR (Fig. 6). (*3*) Libraries 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 (×10⁶)	Real size (×10⁶)	% 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, . . . , *N*. For each *n* sequences, entropy was calculated according to Eq. **1**. Construction of a

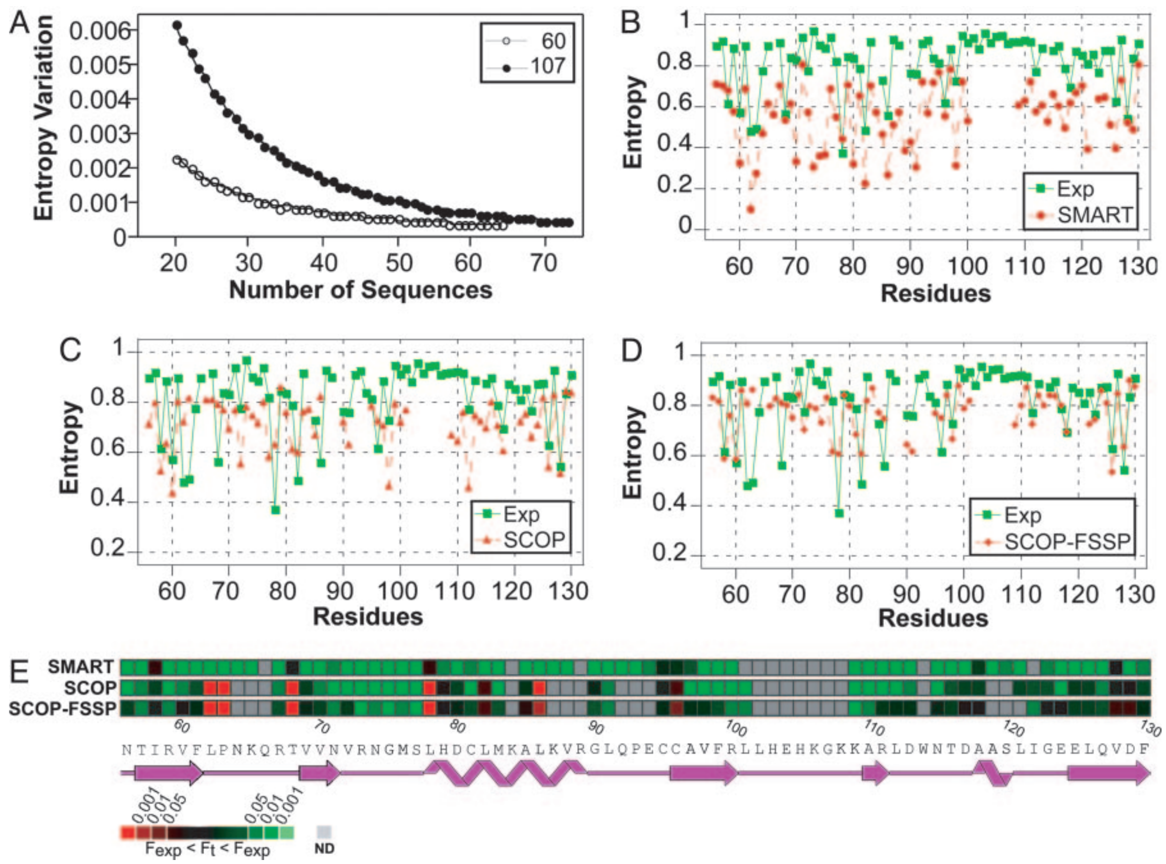

Fig. 2. Validation of experimental data set and comparison of entropy profiles. (*A*) The variation in mean entropy calculated for two successive values of *n* (*n* and *n* + 1) is plotted for V60 and G107 representing, respectively, ''low'' and ''high'' entropy position. Next, the entropy profile obtained 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

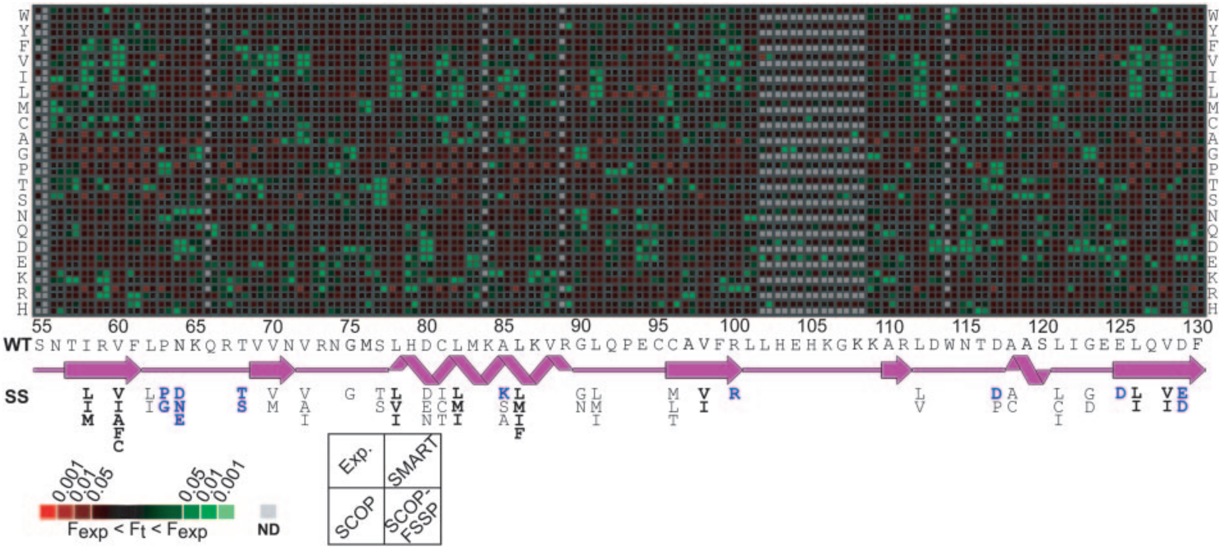

Fig. 3. Amino acid selections. The z scores of aa selections for experimental libraries and the three natural sequence MSAs are represented in a color-coded matrix after the color scale shown. Under each residue, displayed left to right on top of the matrix, there is a column of 20 cells corresponding 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

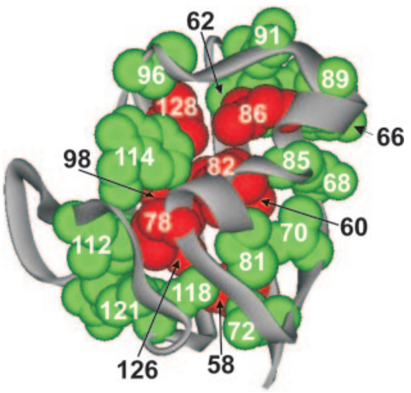

Fig. 4. Hierarchy in the hydrophobic core. Structural disposition of residues that constitute the inner (red) and outer core (green) of Raf RBD. Note that three residues for which insufficient experimental data are available likely belong to the outer core as we define it (Q66, R89, and W114).

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).

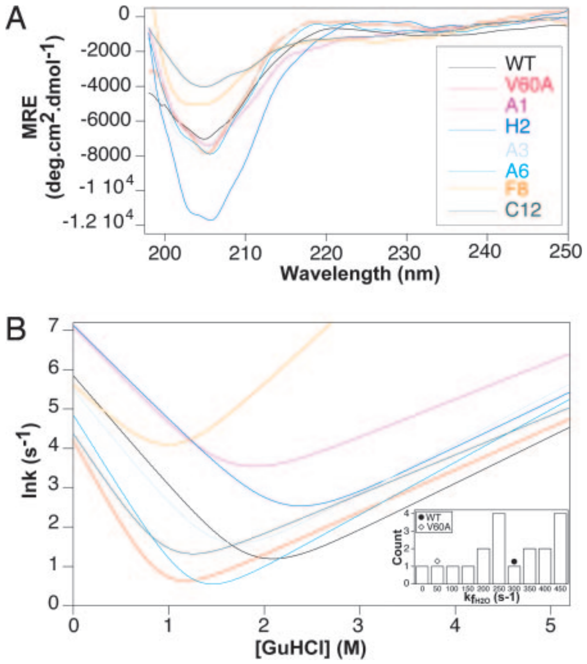

Fig. 5. Clones selected display characteristics of folded proteins. (*A*) Circular dichroism spectra for five clones from five libraries interspersed in the Raf RBD sequence. (*B*) Chevron curves obtained from folding and unfolding experiments in GuHCl for the clones presented in *A*. *Inset* shows the 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} $\cdot$bp$^{-1}\cdot$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.