Comprehensive Analysis of Loops at Protein-Protein Interfaces for Macrocycle Design

Jason Gavenonis*, Bradley A. Sheneman*, Timothy R. Siegert*, Matthew R. Eshelman and Joshua A. Kritzer

*These authors contributed equally to this work.

Department of Chemistry, Tufts University, 62 Talbot Avenue, Medford MA 02155

joshua.kritzer@tufts.edu

ABSTRACT

Inhibiting protein-protein interactions (PPIs) with synthetic molecules remains a frontier of chemical biology. Many PPIs have been successfully targeted by mimicking α-helices at interfaces, but most PPIs are mediated by non-helical, non-strand peptide loops. We sought to comprehensively identify and analyze these loop-mediated PPIs by writing and implementing LoopFinder, a customizable program that can identify loop-mediated PPIs within all protein-protein complexes in the Protein Data Bank. Comprehensive analysis of the entire set of 25,005 interface loops revealed common structural motifs and unique features that distinguish loop-mediated PPIs from other PPIs. "Hot loops," named in analogy to protein hot spots, were identified as loops with favorable properties for mimicry using synthetic molecules. The hot loops and their binding partners represent new and promising PPIs for the development of macrocycle and constrained peptide inhibitors.

INTRODUCTION

Specific interactions between proteins are responsible for a wide range of signaling processes in the cell. As a result, targeting protein-protein interactions (PPIs) is a growing field of drug discovery. [1,2] Recent work using macrocyclic molecules has demonstrated that such compounds are particularly adept at inhibiting PPIs.[3] Macrocyclic natural products such as polyketides and non-ribosomally-synthesized peptides can have exquisite potency and selectivity, and recent approaches have developed synthetic macrocycles that approach their sophistication. Decades of work with selection techniques such as phage display and RNA display have revealed that cyclization almost universally augments the affinities and selectivities of the selected molecules. Macrocyclic linkers and other conformational constraints can endow even large peptides and proteins with surprising bioavailability, and these effects have been observed in areas as diverse as the optimization of peptide hormones and the engineering of highly disulfide-bonded natural products.[4] Macrocycles are theoretically capable of inhibiting nearly any PPI, but PPIs mediated by short peptide loops provide the most direct starting point for designing macrocyclic inhibitors.

Designing inhibitors is not straightforward, but peptides and peptidomimetics offer the advantage of being able to directly mimic specific secondary structures. β-turns and β-strands are readily mimicked by a diverse collection of small-molecule and peptide scaffolds.[5,6] Numerous strategies are also available for mimicking or structurally stabilizing α-helices, including side-chain-to-side-chain crosslinks, backbone-to-backbone crosslinks, backbone replacements using unnatural residues such as β-amino acids, and non-peptidic scaffolds such as terphenyls or macrocycles.[7] These and other classes of molecules have been used to target important helix-mediated PPIs, such as the p53-MDM2 and Bcl-xL-BH3 interactions. While campaigns focusing on inhibiting helix-mediated PPIs have been successful, a survey of the Protein Data Bank showed that only 26% of interface residues have α-helical secondary structure, with 24% having β-strand secondary structure and the remaining 50% having non-regular secondary structure.[1]

Some systematic methods have been developed to use structures of PPIs to identify druggable pockets and to design potential inhibitors. HotSprint searches for conserved residues at PPIs that satisfy requirements for solvent accessibility.[8] Another approach searches PPI interfaces for regions with maximal changes in solvent-accessible surface area upon complexation, then uses these "anchor residues" as pharmacophores to design inhibitors.[9] The relative accessibility of α-helix mimetics prompted a systematic survey of hot spots located within α-helices at protein-protein interfaces.[10,11] In addition, an algorithm called PeptiDerive was used to search a set of 151 pre-selected PPIs for short segments that contain multiple hot spots, regardless of structure.[12] The EphB4-ephrin B2 interaction was cited as a proof-of-concept result for PeptiDerive, specifically residues 116 to 128 on ephrin B2.[12] A similar sequence discovered via phage display was found to be antagonistic for EphB4 with an IC_{50} of 15 nM, validating the approach.[13] However, this kind of analysis has never been integrated for batch processing of the entire PDB, nor has it been done with customizable parameters that allow loops of interest to be defined by structure.

To explore loop-mediated PPIs and to facilitate the design of macrocyclic inhibitors, we sought to comprehensively identify all known protein complexes that are mediated by short peptide loops. Herein, we describe LoopFinder, an original program for comprehensively searching structural databases for peptide loops at PPIs. We used LoopFinder to identify a set of loops that contribute significantly to binding interactions – in analogy to hot spots, we chose to call these "hot loops". These hot loops identify novel targets for inhibition and provide starting points for the rational design of macrocycles as PPI inhibitors.

RESULTS

Workflow for LoopFinder is depicted in Supplementary Results (Supplementary Figure 1). We acquired 19,657 multi-chain structures from the PDB in August 2013, representing all multi-chain structures with ≤ 4Å resolution and < 90% sequence identity. PDB files were manipulated with a C++ script to remove headers and to define each binary protein-protein interface within multi-protein complexes (for NMR structures, only the first structure in the file was analyzed). These interfaces were then inputted in bulk to LoopFinder, which identified peptide loops at protein-protein interfaces as defined by several parameters. First, loops were limited to segments of 4 to 8 consecutive amino acids, in order to conform to molecular mass ranges typical for useful peptide and macrocycle ligands. Another parameter required at least 80% of residues within the loop to reside near the protein-protein interface (having at least one atom within 6.5 Å of the binding partner). We also included a 6.2 Å cutoff between the alpha carbons of the loop termini, to ensure a loop-like conformation and to exclude repeating secondary structures such as β-strands and α-helices. This distance was further restricted to a maximum of 4.67 Å for four-amino-acid loops and 5.83 Å for five-amino-acid loops, in order to eliminate non-loop structures. All of these specific parameters were designed to identify loops that might be amenable to mimicry by small cyclic peptides and other macrocycles. With these parameters, LoopFinder identified 121,086 total loops in 9,388 different structures, including numerous redundancies such as overlapping loops, nested loops, and homologous loops on different chains of homomultimeric protein complexes. These redundant loops were retained for computational alanine scanning because they would not significantly increase the computational burden, and would allow us to select the most critical loops from among them after interaction energies were assigned to each residue.

The complete set of 121,086 loops was then analyzed by computational alanine scanning with PyRosetta v2.012 and Rosetta 3.0 using a modified version of the previously reported scoring function that lacked environment-dependent hydrogen-bonding terms.[14–17] The computational alanine scan produced data that is interpreted as the relative, predicted $\Delta\Delta G_{residue}$ for each residue for that particular PPI. At this point, the loop set was consolidated to eliminate redundant loops, producing a master list of 25,005 interface loops. These interface loops were then sorted by the presence and relative location of hot spot residues, with hot spots defined as $\Delta\Delta G_{residue} \geq 1$ kcal/mol. To generate the set of "hot loops," we identified loops with two consecutive hot spots, loops with at least three hot spots, and loops for which the average $\Delta\Delta G_{residue}$ was greater than 1 kcal/mol (Figure 1). This yielded a set of 1,407 hot loops, covering 1,242 multi-chain PDB structures (Supplementary Data Set 1). This represents only 5.6% of the interface loops, highlighting that this process identified those loops that are most critical for mediating PPIs. Further, for each hot loop, the

total predicted energy associated with the hot loop was compared to the total predicted energy for the corresponding interface. This analysis revealed that 36% of hot loops are responsible more than half of the predicted binding energy for the associated interface, and 67% of hot loops are responsible for more than a quarter of the predicted binding energy (Supplementary Figure 2). Overall, hot loops represent a significant percentage of the total interface energy, making them ideal starting points for identifying novel targets for macrocycles and constrained peptides.

Structural classification of interface loops

The sets of 25,005 interface loops and 1,407 hot loops warrant closer characterization in order to better understand loop-mediated PPIs. First, we analyzed the hot loops with respect to loop structure by identifying secondary structures flanking the loops and canonical turns within the loops. Canonical turn motifs were identified by measuring φ and ψ angles and comparing them with motifs from the PDBeMotif database (see Supplementary Note for definitions).[18] A breakdown of loop structures is shown in Figure 2. 61% of the hot loops possess specific turn motifs with characteristic backbone torsions and intramolecular hydrogen bonds. Loops with one helical turn made up 11% of the hot loops, and these were typically at the N-terminal or C-terminal cap of an α-helix. This shows that the parameters of LoopFinder were defined conservatively enough to exclude regular α-helices. The specific α-turns identified by LoopFinder have very little overlap with previously identified α-helix-mediated PPIs (see below).[10,11]

Unsurprisingly, β-turns are the most common turns, present in 31% of all hot loops. Specific subcategories of β-turns that are prominent in the hot loop set include αβ-motifs and Schellmann loops. We also found that other structural motifs commonly overlap β-turn regions within hot loops (Figure 2). Another common motif within hot loops is a turn-like motif in which serine or threonine makes a side-chain-to-backbone hydrogen bond. There are three classes of such motifs, S/T-turns, S/T-motifs and S/T-staples, which together appear in 16% of all hot loops. β-hairpins, defined as any loop that connects two antiparallel β-strands regardless of the presence of a β-turn, are present in 11% of hot loops. β-hairpins have been thoroughly studied and successfully mimicked using peptides and peptidomimetics.[6] Loops in which aspartate or asparagine form side-chain-to-backbone hydrogen bonds (Asx-turns and motifs) appear in 11% of hot loops. One additional small subset of structured hot loops are β-bulges, which are short breaks within β-strands. Structures of representative hot loops from each of these categories are shown in Figure 2, and demonstrate the wide diversity of loop structures identified by LoopFinder. No immediate correlation was observed among the relative three-dimensional orientations of hot spot residues within each structural class, indicating that, while common structural motifs were observed,

different molecular scaffolds may need to be developed to target different structural classes or even different loops within each structural class.

An important finding from the hot loop set is that canonical turn motifs are not essential for loop-mediated PPIs. The loops categorized as "non-canonical" in Figure 2 have unique structures that are nonetheless still excellent starting points for inhibitor design (Supplementary Figure 3). Interestingly, some of these loops act as N-terminal or C-terminal caps of α-helices. There are a wide variety of torsional angles and intramolecular hydrogen bonds within these loops, giving each a unique topology that may promote high selectivity for their respective binding partners. Cyclization of these loops with flexible or rigidified linkers should produce constrained peptides with unique folded structures. These would represent novel three-dimensional scaffolds for targeting these and other PPIs.

How loops use individual residues to mediate PPIs

To quantify the relative energetic contributions of each amino acid to loop-mediated PPIs, we compiled the average $\Delta\Delta G_{residue}$ for each amino acid for both the interface loop set and the hot loop set (Table 1). The overall trends for both sets are similar, indicating that the hot loops are similar to all interface loops with respect to amino acid usage. The amino acids that have the highest average $\Delta\Delta G_{residue}$ within all interface loops are tryptophan, phenylalanine, histidine, aspartate, tyrosine, leucine, glutamate, isoleucine, and valine. These amino acids span charged, hydrophobic, and aromatic residues, and contain several striking features. Phenylalanine, which is disfavored for PPI hot-spots in general,[19] has a higher average $\Delta\Delta G_{residue}$ than tyrosine, and almost as high as tryptophan. Thus, whatever causes phenylalanine to be disfavored as a hot spot for PPIs in general does not affect loop-mediated PPIs, making phenylalanine as important for loop-mediated PPIs as tryptophan. Another surprising result is that histidine is a major contributor to the binding energy of hot loops, whereas histidine is not observed at higher proportions as a hot-spot residue for PPIs in general.[19] Arginine is not a major contributor to binding energy of loop-mediated PPIs, which is a stark contrast to its major role as one of the most common PPI hot spot residues. Finally, leucine and isoleucine have nearly equal average $\Delta\Delta G_{residue}$ within the interface loop dataset, and are similarly enriched in hot spots within those loops. Thus, while isoleucine is ten times more likely to be a hot spot than leucine within all PPIs,[19] leucine and isoleucine contribute nearly equally to loop-mediated PPIs.

To examine the related question of amino acid abundance within interface loops, we normalized the percent abundance of each amino acid within the interface loop set to the propensity of each amino acid to reside at the protein surface. Then, we broke down these data into change in percent abundance (relative

to surface propensity)[20] for hot spot positions and non-hot spot positions. Overall, these values were similar for the interface loop set and the hot loops; data for all interface loops are shown in Figure 3. Some of the results of this analysis are not particularly surprising. For instance, glycine residues are highly enriched in the interface loop set, which is to be expected for loop regions and for some of the specific loop architectures identified in Figure 2. Another expected result is the prominence of large and hydrophobic amino acids, since they commonly mediate PPIs and have high average $\Delta\Delta G_{residue}$ values within loops. The large overabundance of phenylalanine at hot spots within loops agrees with its high average $\Delta\Delta G_{residue}$, confirming that hot loops commonly use phenylalanine hot spots to recognize protein targets.

Proline might also be expected to be prominent in loops, but it is not over-represented in interface loops or hot loops. This is despite the fact that, when it is present, it (on average) contributes significantly to the binding energy (Table 1). Closer examination of the subset of hot loops containing a proline hot spot (179 loops, 13%) further elucidated the roles of "hot prolines" within these loops. For 70% of these loops, the hot proline is the residue that contributes the most to the interaction, and in 11% of these loops it is the only hot spot. In the majority of loops containing a hot proline, the proline sits at the boundary between the loop and an α-helix or β-strand. This suggests that prolines that act as secondary structure breakers can also play prominent roles in intermolecular interactions.

Within the interface loop set, charged amino acids contribute relatively large $\Delta\Delta G_{residue}$. However, these are also generally prominent on protein surfaces, and are therefore not over-represented within loops (Figure 3). Thus, charged amino acids play similar roles in loop-mediated PPIs as they do in all PPIs. However, loop-mediated PPIs use arginine, aspartic acid, and glutamic acid in equal amounts, while PPIs in general use arginine more often than other charged residues. Strikingly, lysine is among the most abundant amino acids at protein surfaces and one of the most abundant amino acids in flexible loop regions,[19–21] but is greatly under-represented within the non-redundant loop set, both overall and at hot spots. This may be because arginine, aspartate, and glutamate can more readily facilitate hydrogen-bond networks with high cooperativity and stability. The extensive underrepresentation of lysine residues within interface loops, both at hot spots and at non-hot spot positions, distinguishes the interface loops found by LoopFinder from other loops located at the protein surface. This indicates that underrepresentation of lysine (along with the other biases noted above) may be a method for identifying interface loops solely from primary sequence data and secondary structure predictions.

Finally, histidine is over-represented at hot spots within loops (Figure 3), and contributes a very high $\Delta\Delta G_{residue}$ on average (Table 1). This overall analysis does not distinguish whether histidine is contributing via polar, aromatic, or hydrophobic interactions. Visual inspection of "hot histidine" residues in the hot loop set indicates that histidine acts mainly as a hydrogen bond donor and acceptor, making specific polar contacts both within the loop and to the binding partner. Hydrophobic, Van der Waals, or π-stacking interactions involving the imizadole appear to play a less important role for these histidines.

Comparing loop-mediated PPIs to helix-mediated PPIs

Prior work has comprehensively analyzed PPIs mediated by α-helices in order to provide novel starting points for designing PPI inhibitors; these were recently compiled in a web-accessible database called HippDB.[11] To evaluate the degree of overlap between this dataset and the loop-mediated interactions identified by LoopFinder, we cross-referenced the interface loop set to the collection of helical segments from HippDB. We found only 463 complexes were identified by both processes, indicating interfaces made up of surface loops and α-helical regions. Focusing on just the hot loops, we identified only 90 protein structures that were also in HippDB, and only 17 of these contained overlapping sequences between the hot loop and the helix (Supplementary Table 5). Even for these, LoopFinder identified a loop containing at least one additional amino acid outside the helical region for all but one. This reveals that the structural spaces identified by the two methods are essentially orthogonal.

Established and novel targets for inhibitor design

Overall, more than half the hot loops contained consecutive hot spots (Figure 1). Since this subset may be amenable to targeting with small molecules and established β-turn mimetics,[5] we chose instead to examine more closely the set of 364 hot loops not containing consecutive hot spots (Supplementary Data Set 2). This set encompasses a diverse set of loop architectures which display hot spot side chains in wide diversity of three-dimensional orientations that may merit the development of new macrocyclic scaffolds. Included in this subset were several established and emerging drug targets that are discussed further below.

A classic PPI, hGH•hGHbp, contains a hot loop

One of the most thoroughly-studied PPIs has been the interaction between human growth hormone (hGH) and the soluble portion of its receptor, human growth hormone binding protein (hGHbp).[22–24] We

identified two hGH•hGHbp structures among our hot loops. The first consists of native hGH and hGHbp in a 1:2 complex (1HWG), and the second is a mutant interface that was re-optimized by phage display (1AXI).[24] LoopFinder identified two hot loops within these structures, P61-E66 of hGH and I165-M170 of hGHbp. The most critical known hot spot within the hGH loop, R64, was accurately identified by our computational alanine scan, and contributed to the inclusion of hGH P61-E66 among the hot loops (Figure 4a). Likewise, the most critical known hot spots within the hGHbp I165-M170 loop, I165 and W169, were also identified by the computational alanine scan and contributed to its inclusion among the hot loops. Overall, the ability to compare our results to such a well-understood PPI speaks to the robustness and predictive power of Rosetta-based computational alanine scans and of LoopFinder.

A transcription factor•repressor complex: Nrf2•Keap1

Cellular oxidative stress is associated with many disease states, including inflammation, cardiovascular disease, cancer, and neurodegenerative diseases. A coordinated program of protection from oxidative stressors called the antioxidant response is regulated by the transcription factor Nrf2 (Nuclear factor erythroid-derived-related factor 2). Under normal cellular conditions, Nrf2 remains at low levels through its PPI with the Kelch-like ECH-associated protein 1 (Keap1), which targets Nrf2 for ubiquitin-mediated degradation. The critical role of the Nrf2•Keap1 complex in the cell's antioxidant response makes it a therapeutically relevant target. However, a lack of specificity associated with this strategy proved to be a major drawback. Inhibition of Nrf2 degradation by blocking the Nrf2•Keap1 PPI has the potential to be a more selective method of Nrf2 activation compared to prior work using nonspecific Michael acceptors to modify cysteines on Keap1. [25] LoopFinder identified a six-residue sequence, D77-E82 (DEETGE), from the crystal structure of Keap1 bound to a Nrf2-derived peptide (2FLU; Figure 4b).[26] This β-hairpin loop was previously identified as critical for this PPI, and a 16-residue peptide containing this loop (A69-L85) binds Keap1 with a K_d of 24 nM, retaining much of the affinity of full-length Nrf2 (K_d of 5 to 9 nM).[27] Shorter peptides within this loop have binding affinity for Keap1 in the 100-350 nM range.[27,28] Macrocyclic peptides derived from this hot loop have also been developed – these have IC_{50} values in the 15 nM range.[29] Thus, in the case of Nrf2•Keap1, LoopFinder identified a key β-hairpin loop that directly translated into peptide inhibitors with low nanomolar affinity. This example illustrates the utility of LoopFinder as a resource for identifying hot loops as starting points for developing PPI inhibitors.

A protease•protease inhibitor complex: TACE•TIMP-3

Another hot loop identified by LoopFinder is the complex between tissue inhibitor of metalloproteinases 3 (TIMP-3) and TNF-α converting enzyme (TACE). Numerous peptide and small molecule inhibitors of TACE, many of which are broad-spectrum matrix metalloproteinase inhibitors, have been identified by

academic and industrial groups.[29] All of these molecules have targeted the catalytic site of TACE, most using a hydroxymate moiety to bind the catalytic zinc. The most successful small molecule dropped out of Phase II clinical trials due to concerns about specificity.[29]

TIMP-3 is an extracellular protein that binds TACE with sub-nanomolar affinity, inhibiting proteinase activity. The TIMP-3•TACE interaction has been extensively studied both *in vitro* and *in vivo*. The TIMP-3•TACE interaction is facilitated by three interface epitopes within TIMP-3 that bind TACE (PDB 3CKI).[29] LoopFinder identified one of these epitopes, called the sC-connector loop, as a hot loop for the TIMP-3•TACE interaction (Figure 4c). The hot loop consists of a six-residue stretch from S64 to G69 (SESLCG), in which S64 and L67 are identified as hot spots ($\Delta\Delta G_{residue}$ = 3.62 kcal/mol and 1.73 kcal/mol respectively). Notably, S64 and L67 are unique residues within TIMP-3 compared to TIMP-1, TIMP-2, and TIMP-4, which show drastically reduced binding to TACE.[29] In addition, this region is structurally constrained within TIMP-3 via a disulfide bond from C68 to Cys1. The binding pocket for the hot loop has been identified an "alternative pocket" on TACE that could be as important as the actual catalytic zinc active site,[29] but this loop within TIMP-3 has never been used as a starting point for designing TACE inhibitors. Thus, identification of S64-G69 as a hot loop suggests specific designs for non-zinc-chelating inhibitors of TACE. Such inhibitors would have immense therapeutic potential as highly selective TACE inhibitors that do not bind other ADAMs or matrix metalloproteinases.

An E3 ligase complex: Skp2•Cks1

$p27^{Kip1}$ is a G1-checkpoint protein that directly and indirectly regulates many components of the eukaryotic cell cycle, and enhanced degradation of $p27^{Kip1}$ is associated with many common cancers. $p27^{Kip1}$ is targeted for proteolysis by the ubiquitin E3 ligase SCF^{Skp2}, which is a complex of Skp1, Cul1, Rbx1, and Skp2.[30] Ubiquitination of $p27^{Kip1}$ also requires the binding of an accessory protein, Cdc kinase subunit 1 (Cks1), to the SCF^{Skp2} complex. In crystal structures of the SCF^{Skp2} complex (2ASS) and SCF^{Skp2} bound to $p27^{Kip1}$ (2AST),[31] LoopFinder identified a six-residue loop on the surface of Cks1 comprising M38 to W43 (MSESEW). This loop contains four consecutive hot spot residues from S39 to E42 ($\Delta\Delta G_{residue}$ = 1.11 kcal/mol, 1.93 kcal/mol, 2.07 kcal/mol, and 2.94 kcal/mol respectively) located at the N-terminal cap of an α-helix (Figure 4d). Experimental mutagenesis has shown that S41 is essential for SCF^{Skp2} complex activity, confirming the importance of this hot loop.[32] The S39-E42 hot loop represents a novel starting point for designing SCF^{Skp2} inhibitors using α-turn-mimicking scaffolds. Because prior small-molecule screening efforts yielded only weak inhibitors,[33–35] it is likely that the shallow Skp2•Cks1interface is better targeted by constrained peptides or macrocycles.

A histone acetyltransferase complex: Msl1•Msl3

MOF (males-absent on the first) is a histone acetyltransferase (HAT) that exclusively catalyzes the acetylation of histone 4 lysine 16 (H4K16). Only a handful of HAT inhibitors have been discovered, and all target the catalytic site. MOF requires complexation with three regulatory proteins (MSL1, MSL2 and MSL3) for activity.[36,37] LoopFinder identified two hot loops as critical for the formation of the MSL complex (4DNC).[36] One of these is V575 to P580 (VAFGRP) on MSL1, with Phe577 and Arg579 as hot spots ($\Delta\Delta G_{residue}$ = 2.74 kcal/mol and 4.5 kcal/mol respectively; Figure 4e). This loop is highly conserved across eukaryotes (Supplementary Figure 7) and forms a β-hairpin-like structure that binds a shallow pocket on MSL3 (2Y0N).[37] Co-immunoprecipitation experiments showed that variants of MSL1 with mutations in this loop have substantially lower MSL3 affinity, and that a residue that makes up the binding pocket on MSL3, F484, is essential for recognition of MSL1.[37] The other hot loop in the MOF-MSL1/2/3 complex is H183 to G188 (HIGNYE) of MOF, which binds MSL1 using hot spot residues Asn186 and Glu188 ($\Delta\Delta G_{residue}$ = 1.54 kcal/mol and 3.23 kcal/mol respectively; Supplementary Figure 8). To our knowledge, there are no known inhibitors of any member of the MSL complex. LoopFinder has thus identified a hot loop and corresponding binding pocket that may represent a novel druggable interface for targeting cellular HAT activity.

DISCUSSION

Despite the large interfaces of most PPIs, it has long been shown that residues at the interface do not contribute equally to the binding interaction.[19–21] "Hot spots" have been defined as individual residues that contribute a significant portion of binding free energy to the overall interaction (often, $\Delta\Delta G \geq 1$ kcal/mol).[38,39–41]A computational alanine scanning engine based on Rosetta has been developed to computationally predict PPI interface hot spots, typically in 79% agreement with experimental values.[14,15,39–41] Using this approach, computational alanine scans of protein-protein complexes have yielded a wide range of information about the different roles specific residues play at PPI interfaces. In this work, we extend this methodology to short peptide loops, and identify unique structures and properties of "hot loops" that play key roles in diverse PPIs.

LoopFinder is a useful tool for searching structure databases for peptide loops at protein-protein interfaces. LoopFinder identified all PPI interface loops in the PDB, 25,005 in total. Parameters within LoopFinder can be tailored to perform custom searches within our dataset or the entire PDB for loops of specific size or geometry. Three criteria (average $\Delta\Delta G_{residue}$, presence of three hot spot residues, and presence of two consecutive hot spot residues) were used to identify "hot loops" as those loops that

contribute maximally to the PPI. Other criteria can be readily added in the future. We speculate that hot loops with two consecutive hot spots may be unique loops that could be mimicked by traditional small molecules, while those with broader interaction surfaces represent starting points for the design of constrained peptides and other macrocycles. While the design of constrained peptides and small molecule macrocycles is inherently complex, the identification of many potentially fruitful starting points and targets will accelerate this growing field. The LoopFinder algorithm, the set of 25,005 interface loops, and our culled list of 1,407 hot loops are valuable resources for the development of constrained peptides and macrocycles, since these typically have well-defined constraints and topologies that must be matched precisely for successful inhibitor design.

Acknowledgments

J.G. was supported in part by NIH/NIGMS IRACDA grant K12GM074869. T.R.S. was supported in part by Dept. of Education GAANN grant P200A090303. This work was supported in part by NIH DP20D007303 to J.A.K. The authors thank the Tufts Technology Services for research cluster access and Prof. Rebecca Scheck for helpful conversations.

Author Contributions

B.S. wrote the LoopFinder code. J.G., B.S. and J.A.K. troubleshooted, debugged, and parameterized Loopfinder and Rosetta-based computational alanine scanning. J.G., T.S., M.E. and J.A.K. analyzed and contextualized data. J.G., T.S., and J.A.K. produced figures and tables and wrote the paper.

Competing Financial Interests

The authors declare no competing financial interests.

Figure 1. Identification of hot loops. Hot loops are identified as those loops that satisfy one or more of three criteria: the average $\Delta\Delta G_{residue}$ over the entire loop is greater than 1 kcal/mol, the loop has three or more hot spot residues ($\Delta\Delta G_{residue} \geq 1$ kcal/mol), and the loop has two or more consecutive hot spot residues. Representative loops that satisfy each of these criteria are shown within the blue, red and yellow circles (structures from 1AXI, 1GK9, and 1L2U, respectively). Some hot loops satisfy two of these criteria, with representative loops from these categories shown in the purple, orange and green boxes (2QNR, 1GK9 and 2FPF, respectively). In addition, 67 hot loops satisfy all three criteria, an example of which is shown in the gray box to the left (2AST). All structures, rendered in Pymol,[42] show the chain at the interface in blue, the binding partner as a gray surface, the hot loop in green, and hot spots in orange

($\Delta\Delta G_{residue} \geq 1$ kcal/mol) and yellow ($\Delta\Delta G_{residue} \geq 2$ kcal/mol). Representative hot loops display a wide range of loop structures and modes of interaction with the partner surface.

Figure 2. Visualization of different loop structures observed among the hot loops. Representative examples of each type of loop are shown within each circle, including: β-turns (2ZZC), Schellman loops (2OL1), αβ-motifs (2DVT), β-bulges (3GBT), β-hairpins (1T3I), Asx-turns and motifs (1LIA), S/T-turns, motifs and staples (1Y1X), and γ-turns (2IX5). The remaining two categories shown above are α-helical regions identified by their backbone torsional angles (2BM8), and loops lacking canonical structural motifs (3KYH). All structures, rendered in Pymol,[42] show the hot loop in green and hot spots in orange ($\Delta\Delta G_{residue} \geq 1$ kcal/mol) or yellow ($\Delta\Delta G_{residue} \geq 2$ kcal/mol).

Residue	Interface Loops			Hot Loops		
	Avg. $\Delta\Delta G_{residue}$	Percent Abundance	Fold Enrichment in Hot Spots	Avg. $\Delta\Delta G_{residue}$	Percent Abundance	Fold Enrichment in Hot Spots
Trp	0.30	1.3	3.4	1.6	2.1	1.8
Phe	0.28	4.0	3.1	1.3	5.6	1.9
His	0.27	2.5	2.5	1.1	3.7	1.6
Asp	0.21	5.9	1.5	1.0	6.9	1.4
Tyr	0.17	3.7	2.5	1.1	4.7	1.6
Leu	0.15	7.7	1.4	0.83	7.6	1.3
Glu	0.15	5.9	1.3	0.86	7.1	1.2
Ile	0.14	4.2	1.5	0.95	4.2	1.3
Val	0.12	5.2	0.73	0.74	4.8	0.82
Ser	0.09	6.8	0.61	0.59	6.8	0.75
Pro	0.09	4.7	0.89	1.3	5.3	1.1
Thr	0.09	5.6	0.69	0.75	5.4	0.86
Asn	0.08	4.9	0.74	0.56	4.8	0.68
Arg	0.07	5.3	1.4	0.83	6.3	1.2
Lys	0.03	4.8	0.67	0.50	4.0	0.82
Ala	0.01	8.0	0.13	0.30	5.4	0.22
Met	0.01	1.9	0.63	0.55	1.4	0.70
Gly	0.00	13	0.00	0.00	9.9	0.00
Gln	-0.02	3.5	0.49	0.53	2.8	0.71
Cys	-0.03	1.5	0.19	0.37	1.1	0.33

Table 1. Amino acid bias within all 25,005 interface loops and within the 1,407 hot loops. The average $\Delta\Delta G_{residue}$, percent abundance, and fold enrichment for hot spots compared to non-hot-spot positions were calculated for each amino acid within each loop set. See Supplementary Tables 1 and 2 for complete data.

Figure 3. Interface loops use a unique set of amino acids to recognize their binding partners. The percent abundances of each amino acid were normalized relative to propensity to reside on a protein surface.[20] These normalized values were further broken down into all residues (blue), hot spot residues (red) and non-hot spot residues (green).

Figure 4. Established and novel targets for inhibitor design. a) LoopFinder identified a hot loop on the surface of hGH that is known to be essential for binding of hGHbp (1HWG).[24] b) Hot loop within the transcription factor Nrf2 that binds its repressor, Keap1 (2FLU).[26] c) The sC-connector loop of TIMP-3 is a hot loop that binds to the S2 pocket of TACE (3CKI).[43] d) The interaction between Skp2 and Cks1 is essential for the formation of the SCF^{Skp2} complex and its ubiquitin E3 ligase activity (2AST).[31] e) Inhibition of the histone acetyltransferase (HAT) MSL complex is a novel target identified by LoopFinder (2Y0N).[37] All structures, rendered in Pymol,[42] show the hot loop in green, and hot spots in orange ($\Delta\Delta G_{residue} \geq 1$ kcal/mol) or yellow ($\Delta\Delta G_{residue} \geq 2$ kcal/mol).

Online Methods

Protein-protein structures were downloaded from the PDB (August, 2013) using the advanced search function. Structures with > 4 Å resolution or with > 90% similarity were excluded. The resultant structures were than analyzed with LoopFinder, a program written in C++. For all PDB files, the headers were first removed. For NMR structures, only the first structure in the file was considered. For multi-chain biological assemblies, PDB files were split into new files containing each possible binary interface.

Interface residues were identified as any amino acid containing at least one heavy atom within 6.5 Å of another heavy atom on the partner chain. Loops were then identified using the following criteria. 1) Loops contain 4-8 consecutive amino acids, a length suitable for incorporation into a synthetically feasible cyclic peptide. 2) The Cα-Cα distance of the loop termini is limited to a maximum of 6.2 Å for loops of 6-8 amino acids (4.67 Å for 4-amino-acid loops; 5.83 Å for 5-amino-acid loops), to exclude extended secondary structure elements such as alpha helices and beta strands and to ensure that the N- and C-termini of the loop are in relative positions amenable to cyclization. 3) Loops must contain at least 80% interface residues (at least one heavy atom within 6.5 Å of another heavy atom on the partner chain).

Interface loops were then subjected to a computational alanine scan using PyRosetta v2.012 and Rosetta 3.0. The PyRosetta alanine scanning script originally developed by the Gray lab (http://graylab.jhu.edu/pyrosetta/downloads/scripts/demos/D090_Ala_scan.py) was modified to implement a modified score function and to be run in a parallel manner on Tufts's research cluster. The score function was parameterized to match previously reported general computational alanine scanning protocols,[14] but without environment-dependent hydrogen-bonding terms:

```
ddG_scorefxn = create_score_function_ws_patch('standard','score12')
ddG_scorefxn.set_weight(fa_atr, 0.44)
ddG_scorefxn.set_weight(fa_rep, 0.07)
ddG_scorefxn.set_weight(fa_sol, 0.32)
ddG_scorefxn.set_weight(hbond_bb_sc, 0.49)
```

The results of the alanine scans were then combined with the loop data from LoopFinder using a Python script. Sequence data for each loop was combined with the computational alanine scan results to derive a set of hot loops. ΔΔG values that exceeded 4.5 kcal/mol were reduced to 4.5 prior to determining $\Delta\Delta G_{loop,avg.}$, $\Delta\Delta G_{loop,sum}$, or $\Delta\Delta G_{interface}$ based on previous limits set for computation alanine scan values and to avoid biasing the data set in favor of loops with a single hot spot.[14,44] These hot spots represented 8.2% of the hot spots found in interface loops, and were further enriched in the set of hot loops (14.8%).

A web-based interface for LoopFinder is currently under construction. In the meantime, LoopFinder is freely available for use. Requests for binary files or code can be sent to joshua.kritzer@tufts.edu.

References

1. Guharoy, M. & Chakrabarti, P. Secondary structure based analysis and classification of biological interfaces: identification of binding motifs in protein–protein interactions. *Bioinformatics* **23,** 1909–1918 (2007).
2. Wells, J. A. & McClendon, C. L. Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. *Nature* **450,** 1001–1009 (2007).
3. Marsault, E. & Peterson, M. L. Macrocycles Are Great Cycles: Applications, Opportunities, and Challenges of Synthetic Macrocycles in Drug Discovery. *J. Med. Chem.* **54,** 1961–2004 (2011).
4. Bock, J. E., Gavenonis, J. & Kritzer, J. A. Getting in Shape: Controlling Peptide Bioactivity and Bioavailability Using Conformational Constraints. *ACS Chem. Biol.* **8,** 488–499 (2012).
5. Vagner, J., Qu, H. & Hruby, V. J. Peptidomimetics, a synthetic tool of drug discovery. *Curr. Opin. Chem. Biol.* **12,** 292–296 (2008).
6. Nowick, J. S. Exploring β-Sheet Structure and Interactions with Chemical Model Systems. *Acc. Chem. Res.* **41,** 1319–1330 (2008).
7. Azzarito, V., Long, K., Murphy, N. S. & Wilson, A. J. Inhibition of [alpha]-helix-mediated protein-protein interactions using designed molecules. *Nat Chem* **5,** 161–173 (2013).
8. Guney, E., Tuncbag, N., Keskin, O. & Gursoy, A. HotSprint: database of computational hot spots in protein interfaces. *Nucleic Acids Res.* **36,** D662–D666 (2008).
9. Koes, D. *et al.* Enabling Large-Scale Design, Synthesis and Validation of Small Molecule Protein-Protein Antagonists. *PLoS ONE* **7,** e32839 (2012).
10. Jochim, A. L. & Arora, P. S. Systematic Analysis of Helical Protein Interfaces Reveals Targets for Synthetic Inhibitors. *ACS Chem. Biol.* **5,** 919–923 (2010).
11. Bergey, C. M., Watkins, A. M. & Arora, P. S. HippDB: A database of readily targeted helical protein-protein interactions. *Bioinformatics* (2013). doi:10.1093/bioinformatics/btt483
12. London, N., Raveh, B., Movshovitz-Attias, D. & Schueler-Furman, O. Can self-inhibitory peptides be derived from the interfaces of globular protein–protein interactions? *Proteins Struct. Funct. Bioinforma.* **78,** 3140–3149 (2010).
13. Chrencik, J. E. *et al.* Structure and Thermodynamic Characterization of the EphB4/Ephrin-B2 Antagonist Peptide Complex Reveals the Determinants for Receptor Specificity. *Structure* **14,** 321–330 (2006).
14. Kortemme, T. & Baker, D. A simple physical model for binding energy hot spots in protein–protein complexes. *Proc. Natl. Acad. Sci.* **99,** 14116–14121 (2002).
15. Kortemme, T., Kim, D. & Baker, D. Computational alanine scanning of protein-protein interfaces. *Sci. STKE Signal Transduct. Knowl. Environ.* **2004,** pl2 (2004).
16. Chaudhury, S., Lyskov, S. & Gray, J. J. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* **26,** 689–691 (2010).
17. Shulman-Peleg, A., Shatsky, M., Nussinov, R. & Wolfson, H. J. Spatial chemical conservation of hot spot interactions in protein-protein complexes. *BMC Biol.* **5,** 43 (2007).
18. Golovin, A. & Henrick, K. MSDmotif: exploring protein sites and motifs. *BMC Bioinformatics* **9,** 312 (2008).
19. Bogan, A. A. & Thorn, K. S. Anatomy of hot spots in protein interfaces. *J. Mol. Biol.* **280,** 1–9 (1998).
20. Janin, J., Miller, S. & Chothia, C. Surface, subunit interfaces and interior of oligomeric proteins. *J. Mol. Biol.* **204,** 155–164 (1988).
21. Tsai, C.-J., Lin, S. L., Wolfson, H. J. & Nussinov, R. Studies of protein-protein interfaces: A statistical analysis of the hydrophobic effect. *Protein Sci.* **6,** 53–64 (1997).
22. Cunningham, B. C. & Wells, J. A. Rational design of receptor-specific variants of human growth hormone. *Proc. Natl. Acad. Sci.* **88,** 3407–3411 (1991).

23. Cunningham, B. C. & Wells, J. A. Comparison of a Structural and a Functional Epitope. *J. Mol. Biol.* **234,** 554–563 (1993).
24. Sundström, M. *et al.* Crystal Structure of an Antagonist Mutant of Human Growth Hormone, G120R, in Complex with Its Receptor at 2.9 Å Resolution. *J. Biol. Chem.* **271,** 32197–32203 (1996).
25. Hong, D. S. *et al.* A Phase I First-in-Human Trial of Bardoxolone Methyl in Patients with Advanced Solid Tumors and Lymphomas. *Clin. Cancer Res.* **18,** 3396–3406 (2012).
26. Lo, S.-C., Li, X., Henzl, M. T., Beamer, L. J. & Hannink, M. Structure of the Keap1 : Nrf2 interface provides mechanistic insight into Nrf2 signaling. *Embo J.* **25,** 3605–3617 (2006).
27. Chen, Y., Inoyama, D., Kong, A.-N. T., Beamer, L. J. & Hu, L. Kinetic Analyses of Keap1–Nrf2 Interaction and Determination of the Minimal Nrf2 Peptide Sequence Required for Keap1 Binding Using Surface Plasmon Resonance. *Chem. Biol. Drug Des.* **78,** 1014–1021 (2011).
28. Hancock, R., Schaap, M., Pfister, H. & Wells, G. Peptide inhibitors of the Keap1-Nrf2 protein-protein interaction with improved binding and cellular activity. *Org. Biomol. Chem.* **11,** 3553–3557 (2013).
29. Horer, S., Reinert, D., Ostmann, K., Hoevels, Y. & Nar, H. Crystal-contact engineering to obtain a crystal form of the Kelch domain of human Keap1 suitable for ligand-soaking experiments. *Acta Crystallogr. Sect. F* **69,** 592–596 (2013).
30. Chen, Q. *et al.* Targeting the p27 E3 ligase SCFSkp2 results in p27-and Skp2-mediated cell-cycle arrest and activation of autophagy. *Blood* **111,** 4690–4699 (2008).
31. Hao, B. *et al.* Structural basis of the Cks1-dependent recognition of p27(Kip1) by the SCFSkp2 ubipuitin ligase. *Mol. Cell* **20,** 9–19 (2005).
32. Sitry, D. *et al.* Three Different Binding Sites of Cks1 Are Required for p27-Ubiquitin Ligation. *J. Biol. Chem.* **277,** 42233–42240 (2002).
33. Huang, K. & Vassilev, L. T. High-throughput screening for inhibitors of the Cks1-Skp2 interaction. *Methods Enzymol.* **399,** 717–728 (2005).
34. Ungermannova, D. *et al.* High-Throughput Screening AlphaScreen Assay for Identification of Small-Molecule Inhibitors of Ubiquitin E3 Ligase SCFSkp2-Cks1. *J. Biomol. Screen.* **18,** 910–920 (2013).
35. Wu, L. *et al.* Specific Small Molecule Inhibitors of Skp2-Mediated p27 Degradation. *Chem. Biol.* **19,** 1515–1524 (2012).
36. Huang, J. *et al.* Structural insight into the regulation of MOF in the male-specific lethal complex and the non-specific lethal complex. *Cell Res* **22,** 1078–1081 (2012).
37. Kadlec, J. *et al.* Structural basis for MOF and MSL3 recruitment into the dosage compensation complex by MSL1. *Nat Struct Mol Biol* **18,** 142–149 (2011).
38. Clackson, T. & Wells, J. A hot spot of binding energy in a hormone-receptor interface. *Science* **267,** 383–386 (1995).
39. DeLano, W. L. Unraveling hot spots in binding interfaces: progress and challenges. *Curr. Opin. Struct. Biol.* **12,** 14–20 (2002).
40. Gohlke, H., Kiel, C. & Case, D. A. Insights into Protein–Protein Binding by Binding Free Energy Calculation and Free Energy Decomposition for the Ras–Raf and Ras–RalGDS Complexes. *J. Mol. Biol.* **330,** 891–913 (2003).
41. Moreira, I. S., Fernandes, P. A. & Ramos, M. J. Hot spots—A review of the protein–protein interface determinant amino-acid residues. *Proteins Struct. Funct. Bioinforma.* **68,** 803–812 (2007).
42. DeLano, W. L. *The PyMol Molecular Graphic System*. (DeLano Scientific LLC). at <www.delanoscientific.com>
43. Wisniewska, M. *et al.* Structural Determinants of the ADAM Inhibition by TIMP-3: Crystal Structure of the TACE-N-TIMP-3 Complex. *J. Mol. Biol.* **381,** 1307–1319 (2008).
44. Fersht, A. R. *et al.* Hydrogen bonding and biological specificity analysed by protein engineering. *Nature* **314,** 235–238 (1985).

Supplementary Information for

Comprehensive Analysis of Loops at Protein-Protein Interfaces for Macrocycle Design

Jason Gavenonis*, Brad Sheneman*, Timothy R. Siegert*, Matt Eshelman and Joshua Kritzer
*These authors contributed equally to this work

Department of Chemistry
Tufts University
Medford, Massachusetts 02155

SUPPLEMENTARY NOTE

Defining loop structures using PDBeMotif

In order to characterize the types of loops identified as hot loops, the list of 1,407 hot loops were subjected to analysis using the PDBeMotif program.[3] For this process, the PDB file for each identified hot loop was truncated with a short Python script to include only the residues in the hot loop and the three amino acids directly N-terminal and C-terminal to the identified hot loop. This results in an instant analysis of backbone torsional angles and hydrogen bonding patters that exist only within the loop in question. Below is a list of each type of loop as defined by PDBeMotif as well as the requirements used to identify each type of loop.[3] All comments are reproduced from the PDBeMotif definition file. The "nest" and "niche" types of loops were categorized for the purpose of this paper as non-canonical loops due the fact that they lack any organized structural elements such as secondary structure or hydrogen bonding interactions within the loop. In addition to giving analysis of loop structure, PDBeMotif also gives a read out on secondary structure. In this case, if the loop in question was part of an α-helix, they were defined as such in the analysis, making up the final category of structures identified by LoopFinder.

αβ-motif:
A motif of 5 consecutive residues and two H-bonds in which:
-H-bond between CO of residue (i) and NH of residue (i+4)
-H-bond between CO of residue (i) and NH of residue (i+3)
-φ angles of residue (i+1), (i+2) and (i+3) are negative

asx-motif:
A motif of 5 consecutive residues and two H-bonds in which:
-residue (i) is aspartate or asparagine (Asx)
-side-chain O of residue (i) is H-bonded to the main-chain NH of residue (i+2) or (i+3)
-main-chain CO of residue (i) is H-bonded to the main-chain NH of residue (i+3) or (i+4)

Comments: A common feature of the C- and N-termini of α-helices.

asx-turn:
A motif of three consecutive residues and one H-bond in which:
-residue (i) is Aspartate or Asparagine (Asx)
-the side-chain O of residue (i) is H-bonded to the main-chain NH of residue (i+2)
Sub-categories:
Type I:
residue (i): $-140^\circ < \chi 1 - 120^\circ < -20^\circ$ $\quad -90^\circ < \psi + 120^\circ < 40^\circ$

residue (i+1): $-140^\circ < \varphi < -20^\circ$ $-90^\circ < \psi < 40^\circ$

Type I':

Left-handed form of Type I

Type II:

residue (i): $-140^\circ < \chi 1 - 120^\circ < -20^\circ$ $80^\circ < \psi + 120^\circ < 180^\circ$

residue (i+1): $20^\circ < \varphi < 40^\circ$ $-40^\circ < \psi < 90^\circ$

Type II':

Left-handed form of Type II

Comments: about half of the asx-turns are found at the N-termini of α-helices

beta-bulge:

A motif of three residues within a β-sheet in which the main chains of two consecutive residues are H-bonded to that of the third, and in which the dihedral angles are as follows:

residue (i): $-140^\circ < \varphi < -20^\circ$ $-90^\circ < \psi < 40^\circ$

residue (i+1): $-180^\circ < \varphi < -25^\circ$ or $120^\circ < \varphi < 180^\circ$

$40^\circ < \psi < 180^\circ$ or $-180^\circ < \psi < -120^\circ$

Comments: classic β-bulges are mostly found at the edges of the sheets

beta-bulge-loop(5 or 6):

A motif of three residues within a β-sheet consisting of two H-bonds in which:

-the main-chain NH of residue (i) is H-bonded to the main-chain CO of residue (i+4) (Type 5) or residues (i+5) (Type 6)

-the main-chain CO of residue (i) is H-bonded to the main-chain NH of residue (i+3) (Type 5) or residue (i+4) (Type 6)

Type 1 β-bulge loops have an RL nest at residues i+2 and i+3

Type 2 β-bulge loops have an RL nest at residues i+3 and i+4

Comments: β-bulge loops often occur at the loop ends of β-hairpins

Beta-turn:

A motif of four consecutive residues that may contain one H-bond, which, if present, is between the main-chain CO of the first residue and the main-chain NH of the fourth. It is characterized by the dihedral angles of the second and third residues, which are the basis for the sub-categorization:

Sub-categories:

Type I:

residue (i+1): $-140^\circ < \varphi < -20^\circ$ $-90^\circ < \psi < 40^\circ$

residue (i+2): $-140^\circ < \varphi < -20^\circ$ $-90^\circ < \psi < 40^\circ$

Type I':

Left-handed form of Type I

Type II:

residue (i+1): $-140^\circ < \varphi < -20^\circ$ $80^\circ < \psi < 180^\circ$

residue (i+2): $20^\circ < \varphi < 140^\circ$ $-40^\circ < \psi < 90^\circ$

Type II':

Left-handed form of Type II

Comments: The most common of the small protein motifs. The service presents hydrogen bonded motifs only at the distance between the first and last residues C α atoms must be less than 7 Å. Also the dihedral angles were restricted to not fall into the helical region for the first and the last residues or for the middle two residues.

Gamma-turn:

A motif of three consecutive residues i, i+1, i+2 and one H-bond in which:

-the main-chain O of residue (i) is H-bonded to the main-chain NH of residue (i+2)

Sub-categories:

Classic:

residue (i+1): $35^\circ < \varphi < 115^\circ$ $-104^\circ < \psi < -24^\circ$

Inverse:
residue (i+1): -115˚ < φ < -35˚ 24˚ < ψ < 104˚
Comments:

Nest:
A motif of two consecutive residues with dihedral angles as follows:
Sub-categories:
Type RL:
residue (i): -140˚ < φ < -20˚ -90˚ < ψ < 40˚
residue (i+1): 20˚ < φ < 140˚ -40˚ < ψ < 90˚
Type LR:
In LR nests the values for (i) and (i+1) are interchanged
Comments: Nest should not have any prolines. The nest can form a binding site for an anionic group ('the egg'). Nests frequently occur as parts of other motifs such as schellmann loops.

Niche-3l:
A motif of 3 consecutive residues with dihedral angles as follows:
residue (i-1): no limitations
residue (i): -120˚ < φ < -60˚ -50˚ < ψ < 30˚
residue (i+1): -100˚ < φ < -50˚ 110˚ < ψ < 170˚
Comments: This version of the database contains motifs with the amended definition. The criteria used was that the distance between i-1 and i+1 oxygen atoms must be within 4.5Å, both C-O vectors must look towards each other and the residues must not fall into the helical region.

Niche-3r:
A motif of 3 consecutive residues with dihedral angles as follows:
residue (i-1): no limitations
residue (i): -120˚ < φ < -60˚ -50˚ < ψ < 30˚
residue (i+1): -100˚ < φ < -50˚ 110˚ < ψ < 170˚
Comments: This version of the database contains motifs with the amended definition. The criteria used was that the distance between i-1 and i+1 oxygen atoms must be within 4.5Å, both C-O vectors must look towards each other and the residues must not fall into the helical region.

Niche-4l:
A motif of 4 consecutive residues with dihedral angles as follows:
residue (i-1): no limitations
residue (i): -160˚ < φ < -30˚ 90˚ < ψ < 180˚
residue (i+1): 50˚ < φ < 140˚ -40˚ < ψ < 50˚
residue (i+2): -160˚ < φ < -30˚ 90˚ < ψ < 180˚
Comments: According to these definitions, virtually all niche4s incorporate a niche3. This version of the database contains motifs with the amended definition. The criteria used was that the distance between i-1 and i+1 oxygen atoms must be within 4.5Å, both C-O vectors must look towards each other and the residues must not fall into the helical region.

Niche-4r:
A motif of 4 consecutive residues with dihedral angles as follows:
residue (i-1): no limitations
residue (i): -160˚ < φ < -30˚ 90˚ < ψ < 180˚
residue (i+1): -140˚ < φ < -50˚ -50˚ < ψ < 40˚
residue (i+2): -160˚ < φ < -30˚ 90˚ < ψ < 180˚
Comments: According to these definitions, virtually all niche4s incorporate a niche3. This version of the database contains motifs with the amended definition. The criteria used was that

the distance between i-1 and i+1 oxygen atoms must be within 4.5Å, both C-O vectors must look towards each other and the residues must not fall into the helical region.

Schellman-loop-6:

A motif of six consecutive residues (common type) or seven consecutive residues (wide type) that contains two H-bonds in which:

-the main-chain CO of residue (i) is H-bonded to the main-chain NH of residue (i+5) (common type) or residue (i+6) (wide type)

-the main-chain CO of residue (i+1) is H-bonded to the main-chain NH of residue (i+4) (common type) or residue (i+5) (wide type)

Comments: The common Schellman Loop contains an RL Nest at positions (i+3) and (i+4), whereas the wide type contains one at positions (i+4) and (i+5). Schellman Loops are a common feature of the C-termini of α-helices.

ST-motif:

A motif of 5 consecutive residues and two H-bonds in which:

-residue (i) is Serine (S) or Threonine (T)

-side-chain O of residue (i) is H-bonded to the main-chain NH of residue (i+2) or (i+3)

-main-chain CO of residue (i) is H-bonded to the main-chain NH of residue (i+3) or (i+4)

Comments: the ST-motif is analogous to the Asx-motif, with the oxygen of the hydroxyl side-chain replacing that of the acid or amide.

ST-staple:

A motif of four or five consecutive residues and one H-bond in which:

-residue (i) is Serine (S) or Threonine (T)

-the side-chain OH of residue (i) is H-bonded to the main-chain CO of residue (i-3) or (i-4)

-φ angles of residues (i-1), (i-2) and (i-3) are negative

Comments: This is a common feature in the middle part of α-helices. The Ser or Thr side-chain can be regarded as stapling together two adjacent turns of the α-helices.

ST-turn:

A motif of three consecutive residues and one H-bond in which:

-residue (i) is Serine (S) or Threonine (T)

-the side-chain O of residue (i) is H-bonded to the main-chain NH of residue (i+2)

Sub-categories:

Type I:

residue (i+1): $-140^\circ < \chi 1 - 120^\circ < -20^\circ$ $-90^\circ < \psi + 120^\circ < 40^\circ$

residue (i+2): $-140^\circ < \varphi < -20^\circ$ $-90^\circ < \psi < 40^\circ$

Type I':

Left-handed form of Type I

Type II:

residue (i+1): $-140^\circ < \chi 1 - 120^\circ < -20^\circ$ $80^\circ < \psi + 120^\circ < 180^\circ$

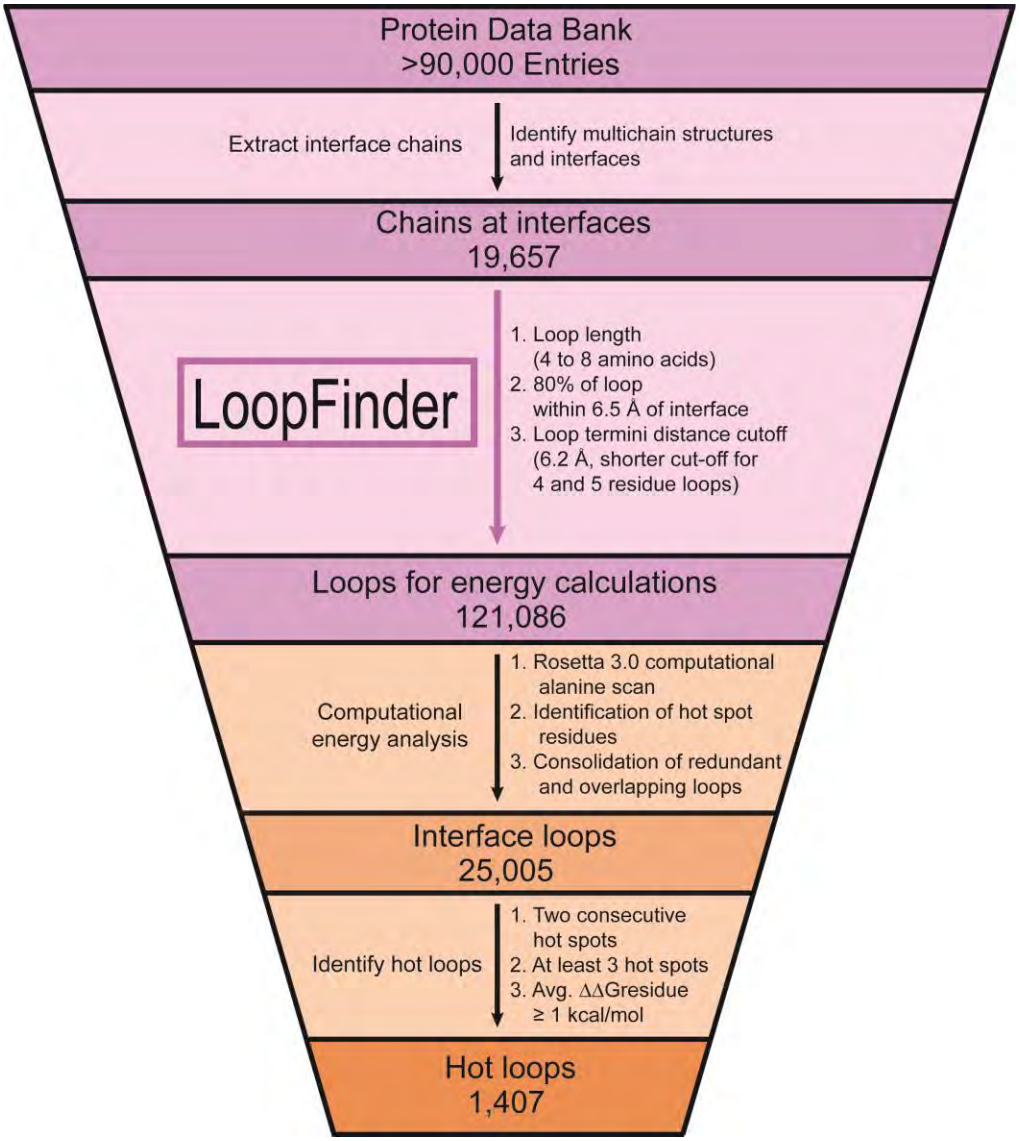
residue (i+2): $20^\circ < \varphi < 140^\circ$ $-40^\circ < \psi < 90^\circ$

Type II':

Left-handed form of Type II

Comments: The ST-turn is analogous to the Asx-turn, with the oxygen of the hydroxyl side-chain replacing that of the acid or amide.

SUPPLEMENTARY RESULTS



Supplementary Figure 1. Workflow for using LoopFinder to identify hot loops at PPI interfaces. LoopFinder was used to analyze all heterogeneous protein-protein complexes in the PDB, identifying just over 121,000 loops at protein-protein interfaces with the given structural parameters. These loop regions were subjected to computational alanine scan mutagenesis using Rosetta to calculate the relative contributions of each residue to the PPI.[14,15] The resulting data were consolidated into a set of 25,005 interface loops, representing a non-redundant set of all loops that mediate PPIs, each with complete computational alanine scan data. The interface loop set was further sorted by key criteria in order to identify those loops containing large proportions of the overall binding energy of the PPI. These comprise the set of 1,407 hot loops.

Supplementary Table 1. The full analysis of hot spots from the total interface loop data set of 25,005 structures. The avg. ΔΔG value for each amino acid, in alphabetical order, is shown in the first column. The total number of each amino acid present in the interface loop data set is shown in the second column followed directly by the percent abundance of each amino acid: $(\frac{\text{number of each amino acid (column 2)}}{\text{total number of amino acids in all loops}} \times 100)$. The number that each amino acid appears as a hot spot, and the percent of each amin o acid that occurs as a hot spot $(\frac{\text{Number of each amino acid that are hot spots}}{\text{number of each amino acid (column 2)}} \times 100)$. In order to identify which amino acids are favored as hot spots, the fold enrichment was calculated:

$$(\frac{\text{number of each amino acid that are hot spots/total number of hot spots for all residues}}{\text{number of each amino acid (column 2)/total number of all amino acids in all loops}} \times 100).$$

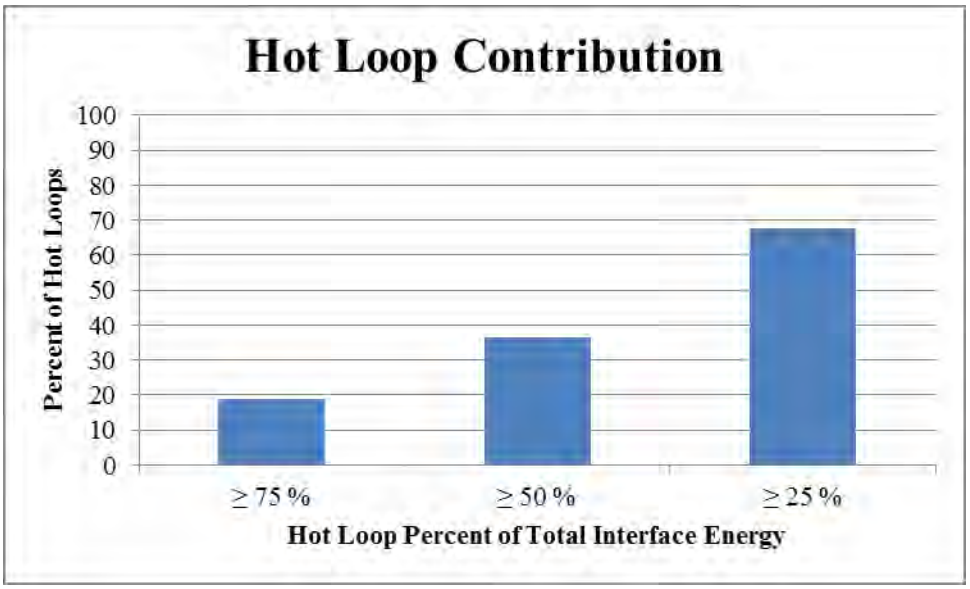
Amino Acid Analysis of Hot Spots in All Loops

Residue	Average $\Delta\Delta G_{residue}$	Total # of amino acids	Percent Abundance	Contributes ≥ 1 kcal/mol		Fold Enrichment in Hot Spots	Contributes ≥ 2 kcal/mol		Fold Enrichment in Hot Spots
				(total)	(%)		(total)	(%)	
Ala	0.01	12341	7.95	98	0.79	0.13	80	0.65	0.36
Arg	0.07	8184	5.27	696	8.50	1.36	184	2.25	1.26
Asn	0.08	7560	4.87	352	4.66	0.74	98	1.30	0.72
Asp	0.21	9169	5.91	881	9.61	1.54	235	2.56	1.43
Cys	-0.03	2291	1.48	27	1.18	0.19	13	0.57	0.32
Gln	-0.02	5378	3.47	165	3.07	0.49	39	0.73	0.41
Glu	0.15	9199	5.93	730	7.94	1.27	217	2.36	1.32
Gly	0.00	19786	12.75	0	0.00	0.00	0	0.00	0.00
His	0.27	3924	2.53	619	15.77	2.52	163	4.15	2.32
Ile	0.14	6502	4.19	592	9.10	1.46	116	1.78	1.00
Leu	0.15	12020	7.75	1048	8.72	1.39	148	1.23	0.69
Lys	0.03	7403	4.77	309	4.17	0.67	64	0.86	0.48
Met	0.01	2955	1.90	117	3.96	0.63	31	1.05	0.59
Phe	0.28	6275	4.04	1201	19.14	3.06	341	5.43	3.04
Pro	0.09	7245	4.67	405	5.59	0.89	230	3.17	1.77
Ser	0.09	10503	6.77	398	3.79	0.61	117	1.11	0.62
Thr	0.09	8639	5.57	372	4.31	0.69	123	1.42	0.80
Trp	0.30	1980	1.28	423	21.36	3.42	221	11.16	6.24
Tyr	0.17	5714	3.68	904	15.82	2.53	223	3.90	2.18
Val	0.12	8092	5.22	367	4.54	0.73	134	1.66	0.93

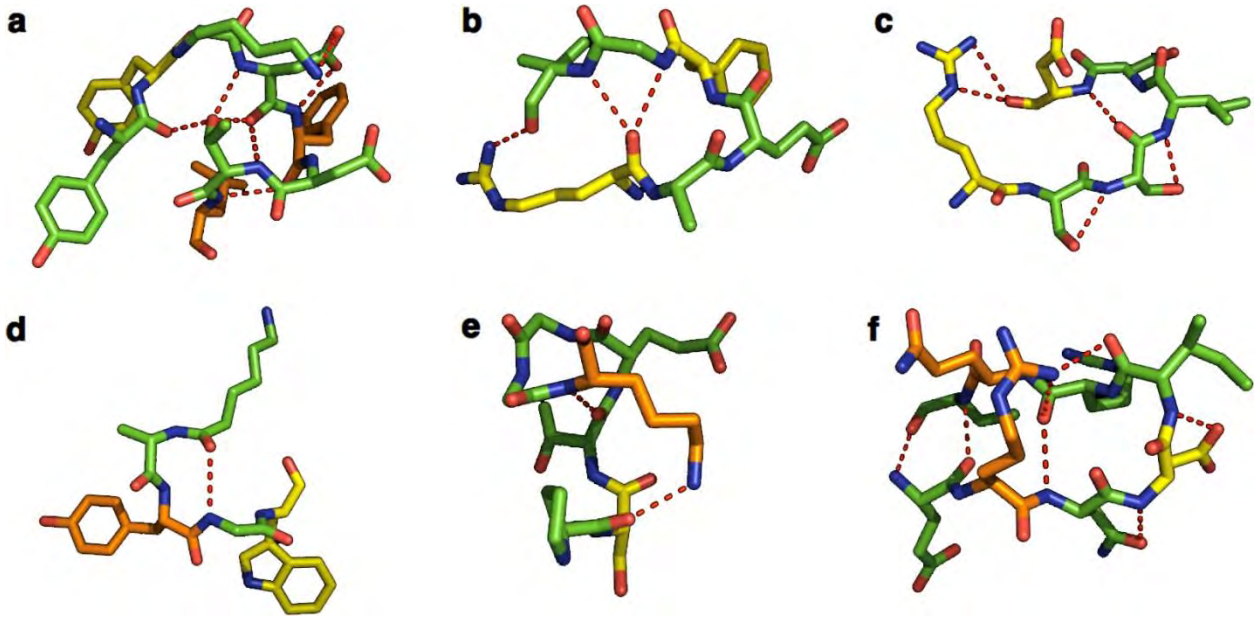
Supplementary Table 2. Identical analysis of hot spots conducted for the data set of hot loops only. Upon comparison to Supplementary Table 1, conclusions can be made as to which amino acid residues are most likely to be contained within a hot loop.

Amino Acid Analysis of Hot Spots in All Loops

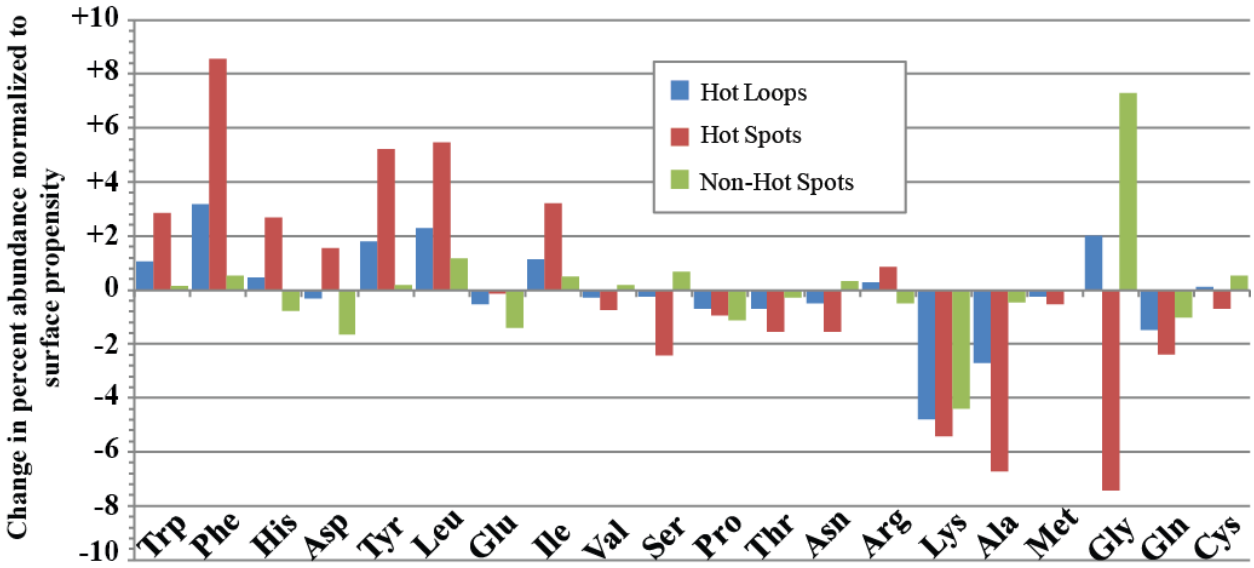
Residue	Average $\Delta\Delta G_{residue}$	Total # of amino acids	Percent Abundance	Contributes ≥ 1 kcal/mol		Fold Enrichment in Hot Spots	Contributes ≥ 2 kcal/mol		Fold Enrichment in Hot Spots
				(total)	(%)		(total)	(%)	
Ala	0.30	468	5.39	36	7.69	0.22	32	6.84	0.54
Arg	0.83	545	6.28	234	42.94	1.23	67	12.29	0.97
Asn	0.56	418	4.82	99	23.68	0.68	40	9.57	0.76
Asp	0.97	597	6.88	284	47.57	1.37	80	13.40	1.06
Cys	0.37	97	1.12	11	11.34	0.33	7	7.22	0.6
Gln	0.53	246	2.84	61	24.80	0.71	21	8.54	0.68
Glu	0.86	613	7.06	262	42.74	1.23	87	14.19	1.12
Gly	0.00	860	9.91	0	0.00	0.00	0	0.00	0.00
His	1.07	319	3.68	181	56.74	1.63	60	18.81	1.49
Ile	0.95	368	4.24	165	44.84	1.29	49	13.32	1.05
Leu	0.83	659	7.59	294	44.61	1.28	65	9.86	0.78
Lys	0.50	348	4.01	99	28.45	0.82	27	7.76	0.61
Met	0.55	119	1.37	29	24.37	0.70	13	10.92	0.86
Phe	1.29	484	5.58	318	65.70	1.89	113	23.35	1.85
Pro	1.27	459	5.29	183	39.87	1.15	125	27.23	2.16
Ser	0.59	594	6.85	155	26.09	0.75	54	9.09	0.72
Thr	0.75	470	5.42	140	29.79	0.86	52	11.06	0.88
Trp	1.55	186	2.14	116	62.37	1.79	78	41.94	3.32
Tyr	1.11	408	4.70	233	57.11	1.64	70	17.16	1.36
Val	0.74	419	4.83	120	28.64	0.82	56	13.37	1.06



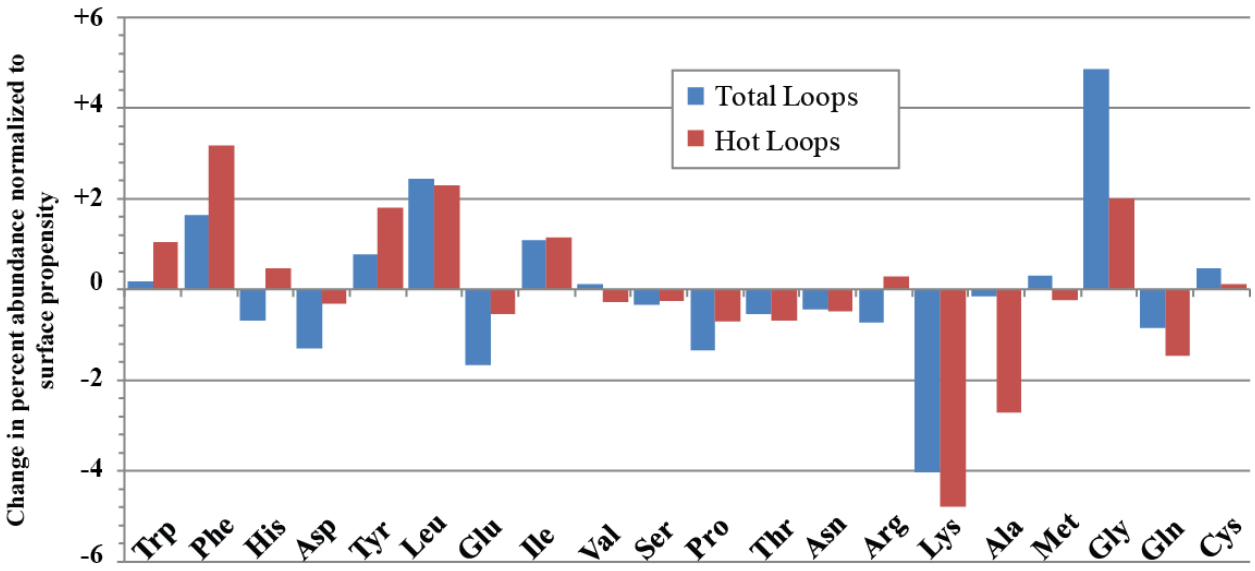
Supplementary Figure 2. Relative contributions of hot loops compared to total interface energy. The contribution of hot loops to the interface energy calculated for the entire chain was analyzed. The resulting contributions show that, on the whole, the hot loops identified by LoopFinder comprise a significant percentage of the total interface energy.



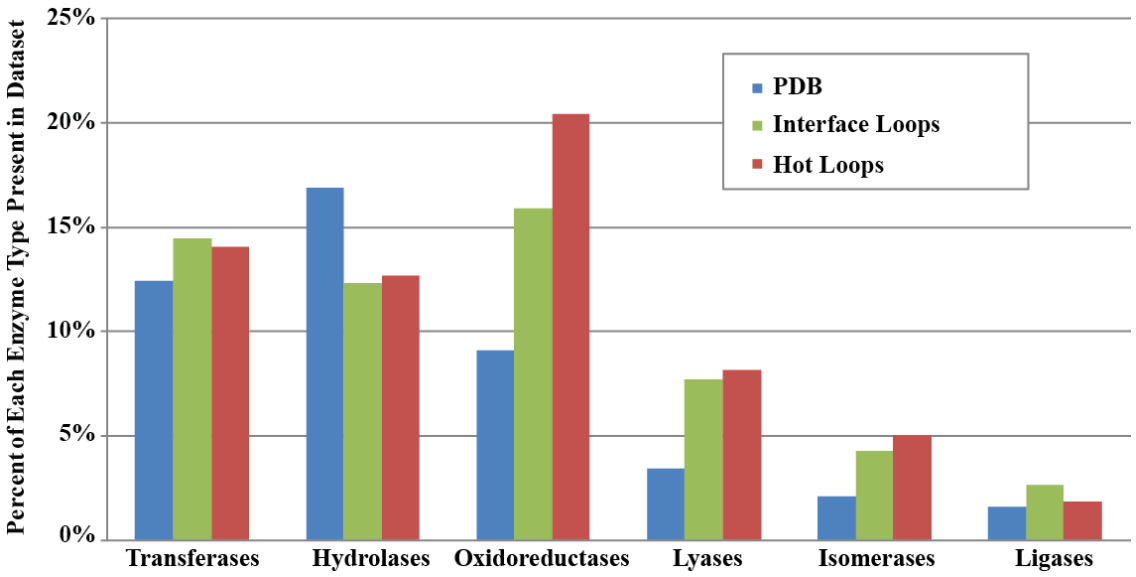
Supplementary Figure 3. Selected hot loops containing unusual turn motifs. a) An interface loop from *Borrelia burgdorferi* BbCRASP-1 with three non-consecutive hydrophobic hot spots constrained by an (*i*,*i*+1) salt bridge and a central threonine with an (*i*,*i*-3;*i*,*i*-6) ST-staple (1W33). b) An interface loop from the *Mycobacterium tuberculosis* toxin-antitoxin complex RelBE2 with a bidentate (*i*,*i*+4;*i*,*i*+5) main-chain hydrogen bond at the C-terminal end of an alpha helix (3G5O). c) An interface loop that contains a β-turn with non-standard backbone torsional angles and two (*i*,*i*+1) side-chain-to-backbone hydrogen bonds (3OA8); d) An interface loop from *Plasmodium falciparium* thioredoxin reductase that contains a β-turn with non-standard backbone torsional angles at the C-terminal end of an α-helix (4B1B). e) An interface loop from potassium ion channel Trek2 with a side-chain-to backbone hydrogen bond from lysine at the N-terminal end of an α-helix (4BW5). f) An interface loop from R-spondin-1 with a β-hairpin-like structure containing an (*i*,*i*+4) hydrogen bond and side-chain-to-backbone hydrogen bonds among Arg (*i*,*i*+3;*i*,*i*+4), Asn (*i*,*i*+1), and Asp (*i*,*i*+1), with both charged residues also being hot spots (4KNG). All structures, rendered in Pymol,[46] show the hot loop in green, and hot spots in orange ($\Delta\Delta G_{residue} \geq 1$ kcal/mol) or yellow ($\Delta\Delta G_{residue} \geq 2$ kcal/mol).



Supplementary Figure 4. Amino acid abundance relative to surface propensity for the hot loop set. The percent abundance of each amino acid in the hot loop data (blue), hot spot residues (red) and non-hot spot residues (green) was compared directly to the natural surface percent abundance of each amino acid as identified previously by Janin *et al.* in order to analyze which amino acids are over-represented at surface hot loops that take part in PPIs.[4] This figure is similar to Figure 3 in the main text, but uses only hot loop data instead of the full interface loop set.



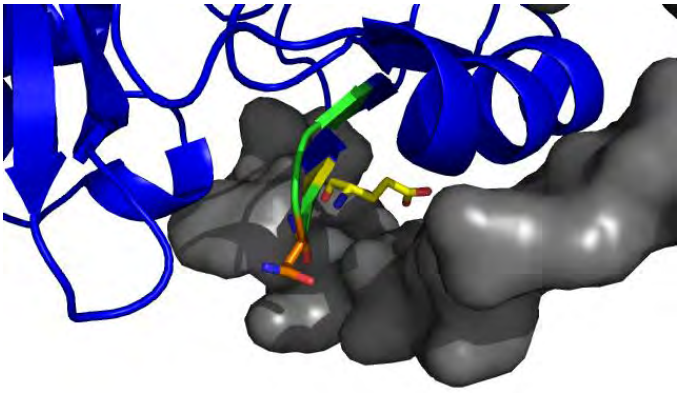
Supplementary Figure 5. Comparison of amino acid abundance between hot loops and all interface loops. A comparison of the amino acid percent abundance, normalized to surface propensity as measured by Janin *et al.*, for total interface loops (blue) and hot loops (red). These data are re-plotted here in one plot to show similarities in amino acid distribution at hot loops compared to the total interface loop set.[4]



Supplementary Figure 6. Loop-mediated PPIs by functional class. Enzymes make up the largest class of proteins with loop-mediated PPIs. Percentages of each enzyme type are shown for the total PDB (blue), the 25,005 interface loops identified by LoopFinder (red), and the 1,407 hot loops (green). The 25,005 interface loops identified by LoopFinder covers 499 different functional classes of proteins from all kingdoms of life, as well as viral proteins. The 1,407 hot loops, by contrast, cover only 132 functions of proteins (Supplementary Data Set 1). The functional classes with the highest number of hot loops are the common classes of enzymes: hydrolases, isomerases, lyases, oxidoreductases, transferases, and ligases. This is not surprising given their prominence in the PDB as a whole. When representation of each enzyme type within the interface loop set and the hot loop set are compared to representation in the input PDB set (as shown here), it is clear that oxidoreductases are over-represented in the interface loop set, and are even more prominent in the hot loop set. Lyases and isomerases show a similar but less prominent trend, and hydrolases show the opposite trend. We conclude that oxidoreductases form loop-mediated PPIs more often than other types of enzymes. Further analysis will illuminate the reason for this over-representation.

```
HUMAN        551   TSFFPEPDDVESLMITPFLPVVAFGR-PLPKLTPQNFELPWLDERSR---
MOUSE        554   TSFFPEPDDVESLLITPFLPVVAFGR-PLPKLAPQNFELPWLDERSR---
CHIMP        289   TSFFPEPDDVESLMITPFLPVVAFGR-PLPKLTPQNFELPWLDERSR---
ZEBRAFISH    427   LSFHAEPEDVEYIMVTPFLPVVAFGS-PLPNLKQQDFDLPWLDERSR---
SALMON       473   SSFYPDTEDVESIIITPFLPVVAFGR-PLPKLTQQNFDLPWQ--RSR---
TETRAODON    364   SSFYPDTEDVETIVITPFLPVVAFGR-PLPKLSQENFELPWLDDRSR---
DROSOPHILA   943   STFYPLPEDIEAIQFVNEVTVQAFGENVVNMEARDDFGVPWVDAIEAPTS
```

Supplementary Figure 7. An alignment of MSL1 across species shows that the region identified as a hot loop is highly conserved across species. In the green box are the loop members that do not contribute significantly to binding. In the yellow box are the hot spot residues identified to contribute significantly to the binding energy.



Supplementary Figure 8. A second hot loop in the MOF complex. The binding interaction between MOF and MSL1 was also identified by LoopFinder (4DNC).[8] The epitope of MOF isolated by LoopFinder ranges from His183 to Glu188 (HIGNYE) with hot spots identified in yellow (Glu188, $\Delta\Delta G_{residue}$ = 3.23 kcal/mol) and orange (Asn186, $\Delta\Delta G_{residue}$ = 1.54kcal/mol). This crystal structure was made using only MSL1, a single component of the MSL complex. It is hypothesized that this same interaction would be present in the full complex structure and allow for the design of an inhibitor aimed at disrupting MOFs ability to bind its activating complex. All structures, rendered in Pymol,[9] show the hot loop in green, and hot spots in orange ($\Delta\Delta G_{residue} \geq 1$ kcal/mol) or yellow ($\Delta\Delta G_{residue} \geq 2$ kcal/mol).

Supplementary Table 3. Comparison to experimental alanine scanning for hGH-hGHbp complexes. Alanine scanning mutagenesis has been extensively done to study the binding of hGH to its hGHbp partner.[5–7] A comparison of mutagenesis results from experimental data is made to the computational data used to identify hot loop for both hGH and hGHbp in two crystal structures isolated in the LoopFinder process. Though the values from the computational alanine scan are not exactly the same as experimental values, the magnitudes are similar. Experimentally, these loops were identified as highly important the for the hGH:hGHbp interaction, confirming our approach for identifying hot loops.

hGH Loop	Experimental $\Delta\Delta G_{residue}$ (kcal/mol)	Computational		hGHbp Loop	Experimental $\Delta\Delta G_{residue}$ (kcal/mol)	Computational	
		1HWG	1AXI			1HWG	1AXI
P61	1.2	0	0	I165	2.0	0.0	4.5
S62	0.1	-0.15	0	Q166	0	0.0	0.1
N63	0.3	4.5	0.4	K167	0	0.3	0.3
R64	1.6	1.9	1.8	G168	n/a	0.0	0.0
E65	-0.5	0.22	0	W169	4.5	-1.1	2.4
E66	n/a	0	0	M170	n/a	0.0	0.0

n/a = no experimental data available

Supplementary Table 4. Checking reproducibility using the online server Robetta. 19 hot loops generated by LoopFinder, each with 3 non-consecutive hotspots, were compared with hot spots calculated independently using the Robetta computational alanine scan server (http://robetta.bakerlab.org/). 57 hot spots were identified using our score function and 54 were identified by Robetta. In total, 44 hot spots (77%) were identified by both methods. Of the hot loops identified, only one (2GE7:A58-63) would be excluded from the hot loop set using this alternate set of ΔΔG values. Underlined residues were identified by both methods; bold by LoopFinder only, italicized by Robetta only.

PDB	CHAIN	FIRST	LAST	AA1	AA2	AA3	AA4	AA5	AA6	AA7	AA8
1IAI	L	HIS91	PHE96	HIS	TYR	SER	**THR**	PRO	PHE		
1J34	A	GLU82	SER87	GLU	TRP	SER	ASP	GLY	**SER**		
1MTP	B	ARG347	LEU353	ARG	*ARG*	ARG	GLY	ALA	*ILE*	LEU	
1XIM	D	SER106	ARG112	SER	ASN	ASP	ARG	SER	VAL	ARG	
2DVT	A	LEU182	TRP187	LEU	LEU	GLY	**PRO**	*THR*	TRP		
2GE7	A	PRO58	CYS63	**PRO**	SER	SER	HIS	ALA	**CYS**		
2IUF	E	ALA33	ALA38	**ALA**	GLY	*GLN*	ARG	GLY	**ALA**		
2OIZ	D	ARG147	LEU154	ARG	PRO	GLY	TYR	GLU	PHE	PHE	LEU
2QLZ	A	LEU84	PHE89	LEU	*THR*	**PRO**	GLU	N/A	PHE		
2QNR	B	LEU257	VAL262	LEU	TYR	PRO	TRP	GLY	**VAL**		
3B7E	A	ALA138	HIS144	ALA	LEU	LEU	ASN	**ASP**	*LYS*	HIS	
3FSL	A	ARG292	SER296	ARG	ARG	ASN	TYR	SER			
3GWA	A	ASP92	SER98	ASP	TYR	*VAL*	LEU	PRO	THR	SER	
3HMU	A	THR90	THR95	THR	PHE	*PHE*	LYS	THR	THR		
3OCD	A	ARG238	ALA244	ARG	GLN	MET	ARG	MET	**PRO**	ALA	
3STH	B	MET305	LYS312	MET	LEU	ASN	ASP	THR	PHE	VAL	LYS
3VTO	C	HIS108	ARG113	HIS	HIS	GLU	GLY	HIS	ARG		
4EHI	A	HIS361	PHE366	HIS	*ILE*	ASP	GLY	GLY	PHE		
4HST	B	LEU37	ASP43	LEU	ALA	**CYS**	ASP	**ARG**	*PHE*	ASP	
Loopfinder Alanine Scan											
PDB	**CHAIN**	**FIRST**	**LAST**	**ΔΔG1**	**ΔΔG2**	**ΔΔG3**	**ΔΔG4**	**ΔΔG5**	**ΔΔG6**	**ΔΔG7**	**ΔΔG8**
1IAI	L	HIS91	PHE96	1.88	0.00	0.00	4.50	0.04	1.46		
1J34	A	GLU82	SER87	0.00	3.24	0.96	1.80	0.00	1.23		
1MTP	B	ARG347	LEU353	2.67	-0.51	2.88	0.00	0.01	0.45	1.54	
1XIM	D	SER106	ARG112	0.00	2.52	0.32	2.02	0.54	0.13	1.90	
2DVT	A	LEU182	TRP187	1.18	0.00	0.00	2.81	0.90	2.93		
2GE7	A	PRO58	CYS63	4.15	0.27	-0.07	1.77	0.00	1.48		
2IUF	E	ALA33	ALA38	1.16	0.00	0.50	1.77	0.00	4.50		
2OIZ	D	ARG147	LEU154	3.12	-0.05	0.00	1.61	-0.87	4.50	0.00	0.00
2QLZ	A	LEU84	PHE89	1.13	0.95	4.50	0.81	0.00	1.10		
2QNR	B	LEU257	VAL262	-0.01	1.30	0.54	2.19	0.00	4.50		
3B7E	A	ALA138	HIS144	0.00	1.77	0.00	0.00	1.58	0.33	3.75	
3FSL	A	ARG292	SER296	4.50	0.59	1.59	0.30	1.16			
3GWA	A	ASP92	SER98	0.72	1.18	-0.90	1.61	0.89	4.50	0.00	
3HMU	A	THR90	THR95	-0.15	3.11	0.00	1.67	0.38	1.28		
3OCD	A	ARG238	ALA244	1.64	0.00	0.00	1.93	0.43	3.83	0.00	
3STH	B	MET305	LYS312	0.00	1.56	0.00	0.00	0.00	2.13	0.00	4.50
3VTO	C	HIS108	ARG113	3.59	0.40	1.51	0.00	2.35	0.00		
4EHI	A	HIS361	PHE366	1.44	0.79	1.44	0.00	0.00	2.45		
4HST	B	LEU37	ASP43	1.33	0.00	4.50	0.27	1.04	0.00	0.00	

Robetta Alanine Scan

PDB	CHAIN	FIRST	LAST	ΔΔG1	ΔΔG2	ΔΔG3	ΔΔG4	ΔΔG5	ΔΔG6	ΔΔG7	ΔΔG8
1IAI	L	HIS91	PHE96	2.32	0.00	0.00	0.29	0.00	1.51		
1J34	A	GLU82	SER87	0.00	5.39	0.53	1.43	0.00	0.50		
1MTP	B	ARG347	LEU353	4.50	4.50	4.02	0.00	0.00	1.75	2.09	
1XIM	D	SER106	ARG112	-0.02	2.61	0.29	4.18	0.55	0.34	4.50	
2DVT	A	LEU182	TRP187	2.07	0.00	0.00	0.00	1.56	4.50		
2GE7	A	PRO58	CYS63	0.00	0.88	0.12	2.47	0.00	-0.05		
2IUF	E	ALA33	ALA38	0.00	0.00	2.21	3.43	0.00	0.00		
2OIZ	D	ARG147	LEU154	2.62	0.00	0.00	4.50	-0.26	2.46	0.00	0.00
2QLZ	A	LEU84	PHE89	2.43	1.37	0.00	0.56	0.00	2.39		
2QNR	B	LEU257	VAL262	0.00	4.37	0.00	4.07	0.00	0.00		
3B7E	A	ALA138	HIS144	0.00	1.60	0.00	0.00	0.75	1.77	3.19	
3FSL	A	ARG292	SER296	1.91	0.40	3.78	0.50	1.28			
3GWA	A	ASP92	SER98	0.90	2.64	1.27	3.31	0.00	3.49	0.00	
3HMU	A	THR90	THR95	-0.23	2.23	2.76	1.16	0.48	3.01		
3OCD	A	ARG238	ALA244	1.34	0.08	0.00	3.60	0.74	0.00	0.00	
3STH	B	MET305	LYS312	0.00	2.66	0.58	0.00	0.00	3.35	0.00	4.50
3VTO	C	HIS108	ARG113	1.79	0.19	2.04	0.00	1.89	0.00		
4EHI	A	HIS361	PHE366	2.22	1.73	3.73	0.00	0.00	3.23		
4HST	B	LEU37	ASP43	2.63	0.00	-0.24	0.16	0.21	1.02	0.00	

Supplementary Table 5. Comparison of LoopFinder results to HippDB results. These show limited overlap between the two databases. Rows highlighted in light gray contain overlapping sequences, with boxed sequences highlighting specific overlapping sequences.

A. PDB IDs for all of the protein complexes identified to have hot loops by LoopFinder and also have helical interface regions as identified by HippDB

B. The interface chains involved in the PPI as identified by LoopFinder, the first letter listed is the chain that contains the interface loop

C. Sequence of hot loop as identified by LoopFinder

D. The interface chains involved in the PPI as identified by HippDB, the first letter listed is the chain that contains the interface helix (reference 10)

E. Sequence of interface helix as identified by HippDB (reference 10)

F. Identical to D, provided for interfaces with multiple interface helices (reference 10)

G. Identical to E, provided for interfaces with multiple interface helices (reference 10)

A. PDB ID	B. Interface Chains	C. LoopFinder Hot Loops	D. Interface Chains	E. HippDB Interface Helices	F. Interface Chains	G. HippDB Interface Helices
1BOU	DC	DEGWG	DA	DLAWHIAQSLIL	DC	IQYLRE
1CPC	AB	ADSQGRFL	BA	RMAACLRDMEIILRY VTYAIFA	LK	RMAACLRDMEIILRYVTYAI FA
1DD4	BA	LTVSEL	CA	TIDEIIEAI	DB	TIDEIIEAIE
			DB	ELAELVKKLEDK		
1DOA	AB	THHCP	BA	SLRKYKEALL		
1XDT	RT	HGERC	BA	LGRLLVV	D C	LGRLLVV
1EEX	ML	DYPLANK	BA	ARPKYQAKSAILHIKET		
1EEX	AL	QRDLKV	BA	ARPKYQAKSAILHIKET		
1EEX	LA	AHGSKD	BA	ARPKYQAKSAILHIKET		
1EFR	AB	HLGESTV	GE	LTLTFNRTRQAVITKELIEIISGA		
1XEY	BA	TIGSSEA	BA	IILRYISYALLA	LK	EIILRYISYALLA
1XFS	AB	LVKNYRP	BA	LGRLLVV	BA	FKLLGNVLVVVLARNF
			DC	GRLLVV		
1GVN	BA	YADDP	DC	NRLNDNLEE	DC	RYETMYAD
1H0H	D	LEAEP	BA	GCQVACKQWH		
1H2K	SA	QGEEL	SA	EELLRAL		

1IWP	AG	FTDGDDT	BA	RPKFMAKAALFHIKE TK	ML	LEKVL
			ML	ERILAIYNAL	ML	FVRESAEVYQQ
1IWP	AL	SHSDIRR	BA	RPKFMAKAALFHIKE TK	ML	LEKVL
			ML	ERILAIYNAL	ML	FVRESAEVYQQ
1IWP	AL	QRDLMV	BA	RPKFMAKAALFHIKE TK	ML	LEKVL
			ML	ERILAIYNAL	ML	FVRESAEVYQQ
1JNR	BA	GYVDYS	CA	VRLQKIMDEY		
1KF6	PO	SAIIA	DC	APVMILLVG		
1KF6	DC	ILLVG	DC	APVMILLVG		
1LIA	LK	LDAFSR	LK	IILRYVSYALLA		
1LQB	CB	YTLKER	CB	LKERCLQVVRSL		
1LTS	ED	MAGKRE	CA	EETQNLSTIYLREYQS KVKRQI	ED	KDTLRITYLT
			FE	KDTLRITYLT	GF	KDTLRITYLT
			HG	KDTLRITYLT		
1M34	JI	RDGFE	DA	KERGRLVDMMTDSHTWL		
1M34	FE	VVCGGF	DA	KERGRLVDMMTDSHTWL		
1MXH	CA	VPLGQ	FB	LRKQR		
1MHY	BD	RWHHPY	GD	GLRKER		
1MHY	DG	WLIEP	GD	GLRKER		
1MTY	BD	KFHGGRPS	HE	LRKQR		
1NVM	AC	VDRETL	CA	YTLMDAADD	GE	YTLMDAADD
1P84	CD	FVFYS	BA	YTKL	DC	RKRLGLKTVIILSSLYLLSIW V...
			GC	ARAYRIIRAHQTELT	HC	VLIPAGIYWYWWKNGNEY NEFL
			FD	EEFFHLQHYLDTAT	GD	AYRIIRAHQTELT
			ID	DTAITSWYENH	ID	WKDVK
1POI	BA	PRSVGD	DC	LRFM		
1PYA	EF	ETKNAYI	CD	DVLDGIVSYDRAET	ED	DVLDGIVSYDRAET
1Q7L	BA	LLHDHDE	D C	DNRYIRA	DC	EAVFLRGVDIYTRLLPALA
1QH8	BA	NRHFKE	DA	ERGRLVDMMLDSHTWL		
1QHH	AB	AKNELL	BA	EKVVSDVYQEYQQRL L	BA	DLIMTTIQLFDR
			CB	AGALAAFRSQLEQWT QL	CB	AQSRLENLDEFLSVTKH
			CB	LIAFLT		
1QHH	BC	YDRKEI	BA	EKVVSDVYQEYQQRL L	BA	DLIMTTIQLFDR
			CB	AGALAAFRSQLEQWT QL	CB	AQSRLENLDEFLSVTKH

			CB	LIAFLT		
1QHH	BC	VIANP	BA	EKVVSDVYQEYQQRLL	BA	DLIMTTIQLFDR
			CB	AGALAAFRSQLEQWTQL	CB	AQSRLENLDEFLSVTKH
			CB	LIAFLT		
1RM6	CA	TQCGFCT	BA	WWRSG		
1RP3	AB	QLIFY	DC	TLSKIAQELS	DC	DEKVVKGLIEFF
			FE	DKVTLSKIAQEL	FE	DEKVVKGLIEFF
			HG	LSKIAQEL	HG	EKKVKELKEKIE
			HG	DEKVVKGLIEFF		
1S5D	ED	LAGKRE	ED	KDTLRIAYLT	HD	IERMKDTLRIAYLT
			FE	DTLRIAYL		
1SDK	BA	AHHFGKEF	D C	LGRLLVV		
1TQY	BA	LWSEG	DC	FTHREFRKLWS	FE	FTHREFRKLWS
1TWF	AH	LTLRDT	KC	HTLGNLIRAE	KC	ALKNACNSIINKLGALKTNFETE
1TZY	FE	KQVHP	BA	IYVYKVLKQV	BA	KAMGIMNSFVNDIFERIAGEAS...
			BA	ELAKHAVSEGTKAVTKYTS	DB	VTYTEH
			DC	PAIRRLARR	DC	EETRGVLKVFLENVIRDAVTYTE
			GC	DIQLARRIR	HF	VTYTEH
			HG	PAIRRLARR	HG	ETRGVLKVFLENVIRDAVTYTE
1ULI	CE	QCRHRG	DC	QMMRGRLRKI	EC	DGENWVEIQQV
			FD	QHEIEQFYYWEAKLLN		
1ULI	CE	EEQAF	DC	QMMRGRLRKI	EC	DGENWVEIQQV
			FD	QHEIEQFYYWEAKLLN		
1UMD	BC	GGHHH	CA	GDWYAGINFAAV	CB	LRQEALL
			DC	LRYR		
1UMD	AC	AHAFGI	CA	GDWYAGINFAAV	CB	LRQEALL
			DC	LRYR		
1YE9	AE	YTEEGI	DA	LREKITHFD	HA	PLLQGRLFSYTD
			FC	PLLQGRLFSYTDTQISR	ED	PLLQGRLFSYTD
			GE	FFAE	HF	FFAE
			LI	REKITHFD	KJ	REKITHFD
			OJ	PLLQGRLFSYTD	NK	ELWEAIE
1YFN	EA	QKMPFW	BA	RPYLLRAFYEWL	DC	PYLLRAFYEWLLD
1YWH	FE	YLWSS	DC	YLWS	HG	YLWS
			NM	YLWS	PO	YLWS
1Z3E	BA	LKRAGI	BA	VRSYNCLKR		
2A6H	ED	VDSKYRL	BA	TLGNPLRRILLS	LK	TLGNPLRRILLS

2AFH	AF	DIVFGG	DA	KERGRLVDMMTDSHTWL		
2AFH	CD	FQKMGI	DA	KERGRLVDMMTDSHTWL		
2B1X	AE	YISEDQ	BA	DSMEMRVLRL	FB	WLYMEAELLD
2B1X	AE	ACRHRG	BA	DSMEMRVLRL	FB	WLYMEAELLD
2B7Y	CD	WDSETGN	D C	LKDV		
2BR2	EB	DYAKKADG	DC	RREIELSKVIREALE	LK	RREIELSKVIREALE
			PO	RREIELSKVIREALE	WX	RREIELSKVIREALE
2BW3	BA	VVRDCR	BA	DCKKEAIEKCAQWVVRD		
2BWE	CD	LRRSGG	BA	EHQLRQLND	CB	EHQLRQLND
			DC	EHQLRQLND	ED	EHQLRQLND
			FE	EHQLRQLN	HG	EHQLRQLN
			IH	EHQLRQLND	KJ	EHQLRQLND
			ML	EHQLRQLND	ON	EHQLRQLN
			QP	EHQLRQLND	RQ	EHQLRQLND
2FM8	AC	DLFALPS	BA	YEILMTI	CB	PALIKQASLDALF
2G38	AB	AADLVS	BA	AAARAWRSLDVEMT AVQRSFNRTL	DC	AAARAWRSLDVEMTAVQR SFNRTL
2GL9	CD	STWNFG	DB	HLVVELMN		
2H5K	BA	RDGAGKY	BA	NELVDY		
2HZS	AB	VFNGSSTG	IB	VDDVLKFTFT	JD	VDDVLKFTFT
			KF	VDDVLKFTFT	LH	VDDVLKFTFT
2J3T	DC	ETDTFK	DC	LASMFHSLFAIGSQ		
2JDI	HG	ASPTQV	IH	YIRYSQICAKAVR		
2P5T	AB	LNPVED	CA	ERYSGYLDGIERMLEI SEKR	DC	HALARNLRSLT
			HD	SYLSTLIRYE		
2XPP	BA	FEIFG	BA	GDRDSLFFEIF		
2QRD	GA	DGETGS	DC	AFLT		
2RF4	AB	IHDAF	BA	LSSSISQLKRIQRDF	DC	LSSSISQLKRIQRDF
2V7Q	FC	HLGESTV	JD	FGKREQAEEERYFRA RAKEQLA...	IH	YIRYSQICAKAVR
2V7Q	AD	SLLLRR	JD	FGKREQAEEERYFRA RAKEQLA...	IH	YIRYSQICAKAVR
2VX8	AB	KSLLG	BA	KFLM		
2WG3	AC	ESRNHV	DB	LDDMEE		
2WNR	BE	FSVEER	FE	SVEISKITAEAL		
2YEV	AB	LSMTPLD	BA	RLEVVWTLIPLAIVFV LFGLTA...	CA	GAALVTLFFYLIL
			CA	DLRFVLFMLLLILLAA GTVALM...	ED	RLEVVWTLIPLAIVFVLFGL TA...
2YFI	EA	CRHRGMRI	FD	QNEIEQFYYREAQLLD	HG	ETMYGRIRKV
			LH	QNEIEQFYYREAQLLD		

2YIU	FA	CTHLG	BA	MDRKQVGFVSVIFLIV LAALLY...	DA	WLHRR
			DA	EVTWIV	ED	MDRKQVGFVSVIFLIVLAAL LY...
2Z5C	CA	NLYYD	CB	AVTHNLYY	FE	AVTHNLYY
2ZC3	AB	YDRNGV	CB	PGDLAEELRR	FE	PGDLAEELRR
3A1G	AB	EKFFP	DC	ERIKELRNL		
3AJV	CB	VDRTGL	CB	WAAAVEVIAG	DB	EIVRAGRL
3AYX	AB	KNPHP	DB	FDEAISE	DC	YMAKLAEQA
			IG	FDEAIS	JH	FDEAISE
			JI	VYMAKLAEQ		
3CF4	AG	AETWQEA	GA	KFYINQVLSAAKNF		
3CIP	AG	ASLSTF	GA	QDESGAAAIFTVQLDDY		
3CR3	CD	SHSPEIA	DC	EIASGLKKLIR		
3DD7	AB	ISRYG	BA	EFASLFDT	DC	AEFASLFDT
3DWL	DA	TDFDGVTF	FD	NGRARLVAETYLSC	KI	ILVRKFMQFL
			KI	IEFMEEVDAEISEMK	KI	NGRARLVAETYLS
3EXE	BA	YYMSGG	BA	TYYM	CA	GQIFEAYNMAAL
			DB	VGAEICARIME	DC	TYYM
			GE	GQIFEAYNMAALW	HF	VGAEICARIME
3EUH	AB	DYYIR	BA	PELVAWARK	BA	RLSFLLAVATLNG
			BA	LGIGITDYYI	BA	EGGDEFHWHRNVYAPLKY
3FXD	BA	YSEII	BA	NTDAVEVLTELNTKV ERA	DC	DNTDAVEVLTELNTKVERA A
3G5O	AB	RAEFGV	BA	LAAVVEFA	DA	WESLQETLYWL
			DC	RESIAEADADIAS	DC	EIRAEF
3H0L	BC	HEGDKT	CA	FQKQLSDILDF	FD	FQKQLSDILDF
			IG	FQKQLSDILDF	LJ	FQKQLSDILDF
			OM	FQKQLSDILDF	RP	FQKQLSDILDF
			US	FQKQLSDILDF	V	FQKQLSDILDF
3HVQ	AC	RIYGFY	CA	ASAEYELE		
3IAM	EG	FREGRY	43	MEAVIYHFKH	DC	MEAVIYHFKH
3K6G	AD	FLKNSG	DA	TLKAAFKTLS	DA	AFAKLDQ
			EB	TLKAAFKTL	EB	AFAKLDQ
			FC	TLKAAFKTL	FC	FAKLDQ
3MM5	ED	QGWIHC	EA	WERFFE		
3MM5	VA	EPPRW	EA	WERFFE		
3O4X	EH	GDYFVF	DA	FFDLKGRLDDIRME	DA	EVFQIILNTV
			DA	EPHFLSILQHL	DA	ARPQYYKLIEECVSQIV
			HE	FWTK		

3OQY	Bb	ERQHM	Aa	AAKERQH	bB	AAKERQH
3P8C	BA	LGPYG	EA	ITSSIKKIADFLNSFDM SCRSR...	ED	SSIKKIADFLNSFDMSCRSRL A...
3P8C	BA	SYHIP	EA	ITSSIKKIADFLNSFDM SCRSR...	ED	SSIKKIADFLNSFDMSCRSRL A...
3RRL	DC	INAGKET	D C	DLVHG		
3RRR	AB	FYQSTCS	BA	IVNKQSCSISNIETVIEF QQK	FB	LHLEGEVNKIKSALLSTNKA VV...
			FD	LHLEGEVNKIKSALLS TNKAVV...	FD	SQVNEKINQSLAFIRKSDEL
			HG	VNKQSCSISNIETVIEF QQK	LH	LEGEVNKIKSALLSTNKAVV SL...
			NH	VSKVLHLEGEV	NH	IKSALLSTNKAVVSLSNGVS VL...
3SDE	AB	CGDGAF	BA	TLAEIAKVE	BA	RWKALIEMEKQQQDQVDR NIKE...
3SQG	AD	GHYGREP	FE	SMDVTAQIHWKRSVGGF		
3U52	BD	YLTRD	DC	MQESAETSFGFCEKR		
3UQY	MT	LGIFR	SM	MSAIITYMVTF		
3UQY	SL	VQSWDDD A	SM	MSAIITYMVTF		
3UQY	LS	GGKNPHPN	SM	MSAIITYMVTF		
3ZWL	BE	HKIFEE	EV	RELLKQWTEYREKIG QEMEKSM	FD	RELLKQWTEYREKIGQEME KSM
4F4O	FD	SCRTA	BA	LGRLLVV	BA	RLLGNVIVVVLARRL
			ED	LGRLLVV	ED	LLGNVIVVVLARRL
			HG	LGRLLVV	HG	LLGNVIVVVLARRL
			KJ	LGRLLVV		
4F6R	BC	KGHYT	BA	LRKLAVNM	CB	AMLERLQEKDKHAEEVRK NK
4FIP	HA	LKKYF	DA	AIVDHL	ED	YFIRHSM
4GD3	SL	VQSWDDD A	SM	SAIITYMVTF		
4GD3	LS	GGKNPHPN	SM	SAIITYMVTF		
4GD3	TM	CIQSGH	SM	SAIITYMVTF		
4GDK	CB	DRLQR	CB	RWKRHISEQLRRRD	CB	FEEIILQYN
			FE	WKRHISEQLRRRDRL	FE	AFEEIILQYN
4GDL	CB	DRLQR	CB	WKRHISEQLRRRDRL	CB	QAFEEIILQYN

Supplementary Table 6. Analysis of all interface loops and the hot loop set with respect to protein function. Annotated function for each protein-protein interaction was identified for the total PDB input set of proteins, the total interface loop set of proteins, and the hot loop set of proteins Only oxidoreductase and lyase enzymes seemed to be more highly represented in the final hot loop data set of proteins compared to the input and total interface loop set. Only categories that had at least one protein of that function in the hot loop dataset are shown and tallied.

A. Functional categories contained in the dataset.

B. Number of each category of protein in the total PDB

C. Number of each category of protein in the total interface loop dataset

D. Number of each category of protein in the hot loop dataset

E. $\frac{\text{Number from B.}}{\text{96,692 (total number of proteins in the PDB)}} \times 100$

F. $\frac{\text{Number from C.}}{\text{25,005 (proteins in the total interface loop set)}} \times 100$

G. $\frac{\text{Number from D.}}{\text{1,407 (proteins in the hot loop set)}} \times 100$

A. CATEGORY	B. # in PDB	C. # in total loop set	D. # in hot loop set	E. % of PDB	F. % of total loop set	G. % of hot loop set
OXIDOREDUCTASE	8786	3980	287	9.09	15.92	20.40
LYASE	3324	1926	115	3.44	7.70	8.17
HYDROLASES	16320	3088	178	16.88	12.35	12.65
ISOMERASE	2026	1070	71	2.10	4.28	5.05
STRUCTURAL GENOMICS	2494	1140	71	2.58	4.56	5.05
TRANSFERASE	12032	3611	198	12.44	14.44	14.07
IMMUNE SYSTEM	2515	716	20	2.60	2.86	1.42
TRANSPORT PROTEIN	2040	311	10	2.11	1.24	0.71
TRANSFERASE/INHIBITOR	1403	83	4	1.45	0.33	0.28
HYDROLASE/HYDROLASE INHIBITOR	2233	344	15	2.31	1.38	1.07
LIGASE	1551	662	26	1.60	2.65	1.85
RNA	946	4	1	0.98	0.02	0.07
TRANSCRIPTION	2779	697	29	2.87	2.79	2.06

RIBOSOME	716	8	1	0.74	0.03	0.07
SIGNALING PROTEIN	2096	353	25	2.17	1.41	1.78
ELECTRON TRANSPORT	1144	144	7	1.18	0.58	0.50
MEMBRANE PROTEIN	1243	253	8	1.29	1.01	0.57
CHAPERONE	955	247	6	0.99	0.99	0.43
DNA BINDING PROTEIN	1261	340	11	1.30	1.36	0.78
TOXIN	890	184	18	0.92	0.74	1.28
FLAVOPROTEIN	89	25	8	0.09	0.10	0.57
METAL BINDING PROTEIN	1044	153	13	1.08	0.61	0.92
RNA BINDING PROTEIN	807	115	6	0.83	0.46	0.43
CELL ADHESION	915	189	7	0.95	0.76	0.50
CELL CYCLE	554	222	14	0.57	0.89	1.00
SUGAR BINDING PROTEIN	887	166	7	0.92	0.66	0.50
VIRAL PROTEIN	1984	421	24	2.05	1.68	1.71
OXIDOREDUCTASE/INHIBITOR	432	34	1	0.45	0.14	0.07
TRANSCRIPTION REGULATOR	405	165	4	0.42	0.66	0.28
PROTEIN BINDING	1249	263	20	1.29	1.05	1.42
APOPTOSIS	399	37	1	0.41	0.15	0.07
UNKNOWN FUNCTION	873	245	17	0.90	0.98	1.21
DE NOVO PROTEIN	342	23	1	0.35	0.09	0.07
ION TRANSPORT	16	11	4	0.02	0.04	0.28
BIOSYNTHETIC PROTEIN	339	136	9	0.35	0.54	0.64
ALLERGEN	112	29	5	0.12	0.12	0.36
HORMONE	371	31	3	0.38	0.12	0.21
OXIDOREDUCTASE/ELECTRON TRANSPORT	53	59	4	0.05	0.24	0.28
OXYGEN TRANSPORT	353	38	5	0.37	0.15	0.36
HORMONE/GROWTH FACTOR	270	23	1	0.28	0.09	0.07
SERINE PROTEASE	132	15	4	0.14	0.06	0.28
PROTEIN TRANSPORT	649	175	7	0.67	0.70	0.50
IMMUNOGLOBULIN	179	39	5	0.19	0.16	0.36
ANTIBIOTIC RESISTANCE	22	13	3	0.02	0.05	0.21
LIPID BINDING PROTEIN	356	45	3	0.37	0.18	0.21
STRUCTURAL PROTEIN	995	208	13	1.03	0.83	0.92
METAL TRANSPORT	378	62	3	0.39	0.25	0.21
LUMINESCENT PROTEIN	221	22	1	0.23	0.09	0.07
GENE REGULATION	410	67	4	0.42	0.27	0.28
TRANSPORT	69	15	3	0.07	0.06	0.21
VIRAL PROTEIN/DE NOVO PROTEIN	2	4	2	0.00	0.02	0.14
LIGHT HARVESTING PROTEIN	3	6	2	0.00	0.02	0.14
PHOTOSYNTHESIS	204	90	4	0.21	0.36	0.28
KETOLISOMERASE	3	21	2	0.00	0.08	0.14
DEHYDROGENASE	16	8	2	0.02	0.03	0.14

PENICILLIN BINDING PROTEIN	14	11	2	0.01	0.04	0.14
CALCIUM BINDING PROTEIN	164	11	1	0.17	0.04	0.07
HEME BINDING PROTEIN	30	19	2	0.03	0.08	0.14
PLANT PROTEIN	252	65	5	0.26	0.26	0.36
LIGASE/INHIBITOR	79	9	2	0.08	0.04	0.14
NEUROPEPTIDE	57	15	2	0.06	0.06	0.14
TRANSLATION	338	92	4	0.35	0.37	0.28
BIOTIN BINDING PROTEIN	144	19	1	0.15	0.08	0.07
HORMONE RECEPTOR	115	8	1	0.12	0.03	0.07
NUCLEAR PROTEIN	95	14	2	0.10	0.06	0.14
LECTIN	166	32	3	0.17	0.13	0.21
TRANSFERASE/HYDROLASE	16	25	1	0.02	0.10	0.07
REPLICATION	210	72	3	0.22	0.29	0.21
MONOOXYGENASE	6	21	1	0.01	0.08	0.07
OXIDOREDUCTASE/IMMUNE SYSTEM	8	21	1	0.01	0.08	0.07
BACTERIAL ANTIBIOTIC RESISTANCE	1	1	1	0.00	0.00	0.07
ADP-RIBOSYLATION	2	1	1	0.00	0.00	0.07
RECEPTOR	120	12	1	0.12	0.05	0.07
ENDOCYTOSIS	69	20	2	0.07	0.08	0.14
IMMUNE SYSTEM/ANTIMICROBIAL PROTEIN	1	2	1	0.00	0.01	0.07
MEMBRANE PROTEIN/CELL ADHESION	1	2	1	0.00	0.01	0.07
POSTSEGREGATIONAL KILLING SYSTEM	1	2	1	0.00	0.01	0.07
TRYPTOPHAN BIOSYNTHESIS	1	2	1	0.00	0.01	0.07
VIRAL PROTEIN/PROTEIN BINDING	5	1	1	0.01	0.00	0.07
NITRITE REDUCTASE	1	3	1	0.00	0.01	0.07
VIRAL PROTEIN/SIGNALING PROTEIN	4	2	1	0.00	0.01	0.07
PROTEIN TURNOVER	3	16	1	0.00	0.06	0.07
MEMBRANE PROTEIN/OXIDOREDUCTASE	2	3	1	0.00	0.01	0.07
PROTEIN BINDING/BLOOD CLOTTING	2	3	1	0.00	0.01	0.07
PROTEIN SYNTHESIS/TRANSFERASE	2	3	1	0.00	0.01	0.07
PROTEIN TRANSPORT/IMMUNE SYSTEM	2	3	1	0.00	0.01	0.07
IMMUNE SYSTEM/CYTOKINE	9	1	1	0.01	0.00	0.07
MEMBRANE TRANSPORT	3	3	1	0.00	0.01	0.07
MOLYBDENUM-IRON PROTEIN	1	4	1	0.00	0.02	0.07
STRUCTURAL PROTEIN/CHAPERONE	4	3	1	0.00	0.01	0.07
DOMAIN SWAPPING	2	4	1	0.00	0.02	0.07
RECOMBINATION	128	34	1	0.13	0.14	0.07
COMPLEMENT REGULATOR	5	3	1	0.01	0.01	0.07
IMMUNOLOGY	1	5	1	0.00	0.02	0.07
PROTEIN BIOSYNTHESIS	1	5	1	0.00	0.02	0.07
HALOPEROXIDASE	14	1	1	0.01	0.00	0.07
ADENOVIRUS	2	5	1	0.00	0.02	0.07

PORIN	10	2	1	0.01	0.01	0.07
IMMUNE SYSTEM/HORMONE RECEPTOR	1	6	1	0.00	0.02	0.07
OXYGENASE	5	4	1	0.01	0.02	0.07
RIBOSOME INHIBITOR	5	4	1	0.01	0.02	0.07
TRANSFERASE/RNA BINDING PROTEIN	5	4	1	0.01	0.02	0.07
OXYGEN TRANSPORT/TRANSPORT PROTEIN	1	7	1	0.00	0.03	0.07
CARBOXY-LYASE	2	6	1	0.00	0.02	0.07
TRANSCRIPTION REGULATION	124	16	1	0.13	0.06	0.07
ANTITOXIN	4	5	1	0.00	0.02	0.07
DNA RECOMBINATION	6	4	1	0.01	0.02	0.07
FORMYLTRANSFERASE	2	7	1	0.00	0.03	0.07
OXIDOREDUCTASE/OXIDOREDUCTASE	2	11	1	0.00	0.04	0.07
HYDROLASE/REPLICATION	3	7	1	0.00	0.03	0.07
LYASE/OXIDOREDUCTASE	3	11	1	0.00	0.04	0.07
TRANSFERASE/RECEPTOR	7	5	1	0.01	0.02	0.07
TRANSCRIPTION/HYDROLASE	4	10	1	0.00	0.04	0.07
HYDROLASE/TRANSFERASE	18	20	1	0.02	0.08	0.07
ACETYLCHOLINE-BINDING PROTEIN/AGONIST TOXIN/ANTITOXIN	17	3	1	0.02	0.01	0.07
OXIDOREDUCTASE/PROTEIN BINDING	7	9	1	0.01	0.04	0.07
GLYCOGEN PHOSPHORYLASE	8	9	1	0.01	0.04	0.07
SIGNALING PROTEIN/PROTEIN BINDING	18	4	1	0.02	0.02	0.07
TRANSCRIPTION/ACTIVATOR	19	4	1	0.02	0.02	0.07
HYDROLASE/HYDROLASE REGULATOR	12	7	1	0.01	0.03	0.07
ENDOCYTOSIS/EXOCYTOSIS	11	8	1	0.01	0.03	0.07
DIOXYGENASE	101	12	1	0.10	0.05	0.07
CELL INVASION	18	13	1	0.02	0.05	0.07
COAGULATION	79	30	1	0.08	0.12	0.07
ANTITUMOR PROTEIN	22	8	1	0.02	0.03	0.07
TRANSCRIPTION REPRESSOR	47	5	1	0.05	0.02	0.07
KINASE	25	8	1	0.03	0.03	0.07
VIRAL PROTEIN/IMMUNE SYSTEM	65	7	1	0.07	0.03	0.07
CIRCADIAN CLOCK PROTEIN	102	29	1	0.11	0.12	0.07
GROWTH FACTOR	38	20	1	0.04	0.08	0.07
TRANSCRIPTION ACTIVATOR	93	16	1	0.10	0.06	0.07
	49	11	1	0.05	0.04	0.07

Supplementary Information References:

(1) Kortemme, T.; Baker, D. *Proc. Natl. Acad. Sci.* **2002**, *99*, 14116–14121.
(2) Fersht, A. R.; Shi, J.-P.; Knill-Jones, J.; Lowe, D. M.; Wilkinson, A. J.; Blow, D. M.; Brick, P.; Carter, P.; Waye, M. M. Y.; Winter, G. *Nature* **1985**, *314*, 235–238.
(3) Golovin, A.; Henrick, K. *BMC Bioinformatics* **2008**, *9*, 312.
(4) Janin, J.; Miller, S.; Chothia, C. *J. Mol. Biol.* **1988**, *204*, 155–164.
(5) Cunningham, B.; Wells, J. *Science* **1989**, *244*, 1081–1085.
(6) Clackson, T.; Wells, J. *Science* **1995**, *267*, 383–386.
(7) Cunningham, B. C.; Wells, J. A. *J. Mol. Biol.* **1993**, *234*, 554–563.
(8) Huang, J.; Wan, B.; Wu, L.; Yang, Y.; Dou, Y.; Lei, M. *Cell Res* **2012**, *22*, 1078–1081.
(9) DeLano, W. L. *The PyMol Molecular Graphic System*; DeLano Scientific LLC: San Francisco, CA.
(10) Bergey, C. M.; Watkins, A. M.; Arora, P. S. *Bioinformatics* **2013,** *29*, 2806–2807.
(11) Kortemme, T.; Kim, D.; Baker, D. *Sci. STKE Signal Transduct. Knowl. Environ.* **2004**, *2004*, pl2.

Supplementary Data Set 1. The entire set of hot loops generated by LoopFinder. These are those loops that meet the requirements previously discussed and outlined in the methods description.

A. PDB ID numbers for each protein complex identified to have a hot loop at the interface
B. Structure title
C. Functional category of the protein that takes part in the interaction
D. The chain on which the hot loop is located
E. The partner that binds to the exposed hot loop
F. The length of the sequence identified as a hot loop, in number of amino acids
G. The number of the first residue in the hot loop
H. The number of the last residue in the hot loop
I. Linker length, the distance between the N and C terminus of the hot loop peptide, in Angstroms
J. The sequence of the hot loop
K. The calculated $\Delta\Delta G_{residue}$ as calculated using computation alanine methods developed by Kortemme et al. (reference 11)
L. The average energy of the loop, $\Delta\Delta G_{loop,avg.}$ is the average of all values from column K.
M. The sum of all $\Delta\Delta G_{residue}$ values from column K. (reference 11)
N. The hot loops percent contribution to the total interface energy
O. Comments: (*) denotes that the $\Delta\Delta G_{loop,sum}$ calculated for the hot loop is a negative value. Negative values denote the possibility that substitution of alanine may improve binding, as in the prevention of an optimal, buried hydrogen bond (as described in reference 1). (+) denotes that the calculated total interface energy is a negative value.
P. Loop types for as identified using PDBeMotif (reference 3)

Supplementary Data Set 2. The subset of 364 hot loops that do not contain two or more consecutive hot spots. These loops were the main focus for identifying biologically relevant protein complexes that may be best targeted using constrained macrocycles.

A. PDB ID numbers for each protein complex identified to have a hot loop at the interface
B. Structure title
C. Functional category of the protein that takes part in the interaction
D. The chain on which the hot loop is located
E. The partner that binds to the exposed hot loop
F. The length of the sequence identified as a hot loop, in number of amino acids
G. The number of the first residue in the hot loop
H. The number of the last residue in the hot loop
I. The sequence of the hot loop
J. The calculated $\Delta\Delta G_{loop}$ as calculated using computation alanine methods developed by Kortemme et al. (reference 11)
K. Number of hot spots located within the loop