

Restricted sidechain plasticity in the structures of native proteins and complexes

Sarel J. Fleishman,¹ Sagar D. Khare,¹ Nobuyasu Koga,¹ and David Baker^{1,2*}

¹Department of Biochemistry, University of Washington, Seattle, Washington 98195

²Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195

Received 30 November 2010; Revised 1 February 2011; Accepted 2 February 2011

DOI: 10.1002/pro.604

Published online 22 February 2011 proteinscience.org

Abstract: Protein-design methodology can now generate models of protein structures and interfaces with computed energies in the range of those of naturally occurring structures. Comparison of the properties of native structures and complexes to isoenergetic design models can provide insight into the properties of the former that reflect selection pressure for factors beyond the energy of the native state. We report here that sidechains in native structures and interfaces are significantly more constrained than designed interfaces and structures with equal computed binding energy or stability, which may reflect selection against potentially deleterious non-native interactions.

Keywords: protein design; negative design; small-molecule binding; monomer design; Rossmann fold design; aromatic residues; Boltzmann distribution; entropy of binding

Introduction

Protein-design methodology has been used to create new protein structures,^{1,2} redesign protein-protein interactions,^{3,4} and produce new enzymes.^{5,6} Although there is still more to improve in protein-design algorithms,⁷ the field has considerable promise for generating new molecules for use in a wide variety of areas of current interest. By providing a reference point for which all of the inputs are known, protein design can also help to cast into

sharper focus the critical evolved properties of native biomolecules. For example, studies of the kinetics of folding of designed proteins have helped distinguish the features of folding that are simple consequences of having a unique folded state from those optimized by natural selection.⁸

To gain insight into the properties of native structures and interfaces that result from more than simple selection for very low-energy native states, we used computational-design methodology to generate designed structures, protein-protein interfaces, and protein-ligand interfaces. As shown in Figure 1(a), designed protein-protein complexes have computed binding energies in the same range as those of native ones. Though the errors in computed energies are likely to be at least several kcal/mol, the computations do capture overall trends reasonably well; for example, the correlation between experimentally

Additional Supporting Information may be found in the online version of this article.

Grant sponsors: Defense Threat Reduction Agency, Defense Advanced Research Projects Agency, HHMI.

*Correspondence to: David Baker, Department of Biochemistry, University of Washington, Seattle, WA 98195.
E-mail: dabaker@uw.edu

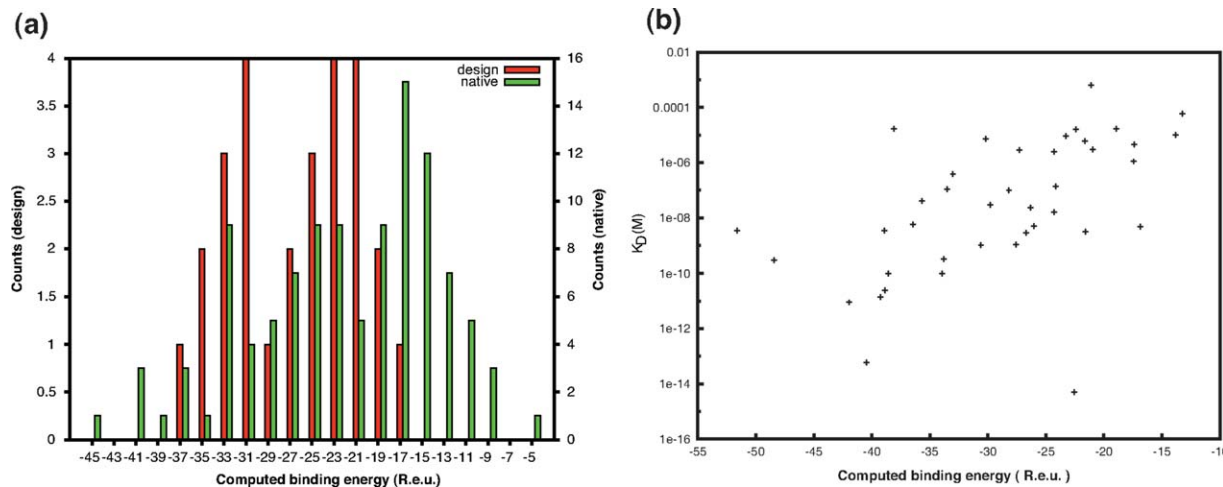


Figure 1. (a) Computed binding energies for designed and native protein-protein interfaces. (b) Experimental versus computed binding energies in natural protein-protein complexes (Pearson correlation coefficient $r = 0.53$). Binding data from Ref. 9.

measured binding affinities and computed binding energies for protein-protein complexes [Fig. 1(b)]. Thus, although the errors in the computed energy of any one design may well be sufficiently large for the designed protein to not fold or the designed interface to not form, the distribution of interactions overall in the native and designed complexes are likely to be similar.

With the initial aim of estimating the sidechain-entropy loss upon protein-complex formation, we developed a simple measure of the Boltzmann weight of sidechain conformations in native and designed interfaces. For protein interfaces, the method first separates the complex, and for each residue that makes an appreciable contribution to binding (see Methods section), iterates over all of its rotameric states as defined in the Dunbrack library of backbone-dependent rotamers,¹⁰ excluding rotamers that are predicted to clash with protein mainchain or C β atoms. For each rotamer placement, all residues within a 6 Å shell are repacked and minimized. The energy E of each such state is then evaluated using the Rosetta all-atom energy function,¹¹ which is dominated by van der Waals, hydrogen bonding, and solvation terms. The probability of rotamer i , p_i is then computed assuming a Boltzmann distribution:

$$p_i = \frac{\exp(-E_i/k_B T)}{\sum_s \exp(-E_s/k_B T)} \quad (1)$$

where s is the rotameric state, k_B is the Boltzmann constant, and T is the absolute temperature. E is the energy of the states.

We estimated the probabilities in the unbound state of aromatic sidechain conformations observed in bound complexes and designed binders according to Eq. (1) [Fig. 2(a,b)]. Both designs and native com-

plexes have low-probability sidechains at the interface that contribute significantly to the binding energy. However, very high-probability conformations ($p > 50\%$) are almost exclusively found in natives, whereas designs are over-represented in the very low-probability bin ($p < 5\%$). The trend extends to monomeric structures: we observe higher probabilities for aromatic residues in native conformations than in designs [Fig. 2(c)]. Thus, although the overall computed energies of the designed structures and complexes are similar to those of native structures, aromatic sidechains in the former are considerably less restricted in conformation than those in the latter. We found no correlation between the computed stability of individual residues in the cores of designed and native monomers and their conformational probability (not shown). The lack of correlation demonstrates that the conformational restrictions we observe in native proteins compared to designs are not a simple consequence of natural proteins having more stable cores than designs.

What is the origin of these differences in sidechain-conformational restriction? The answer may lie in the fact that most current design methods, including those used to generate the structures and complexes in this study, focus on optimizing the energy of the desired target structure or complex. Such “positive design” strategies suffer from the flaw that even though the energy of the target structure is optimized to be extremely low, the target conformation or complex might be high in energy relative to alternative conformations, and hence could be poorly populated.¹² Negative-design methods have been developed that contain explicit bias for sequence substitutions that favor a target structure versus a set of alternative structures,^{13,14} but the latter have to be explicitly modeled and thus for computational tractability the set must not be too large.

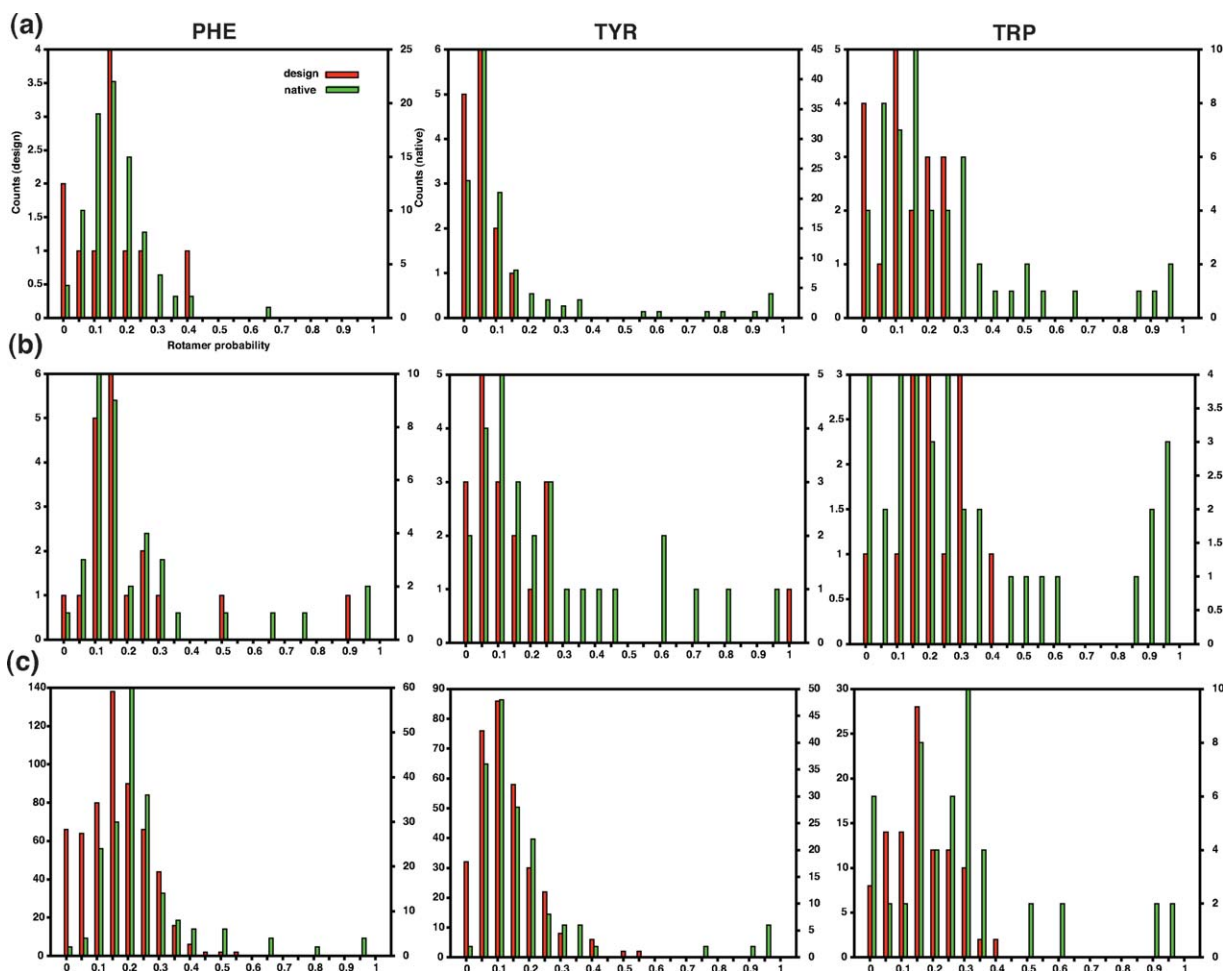


Figure 2. The probabilities in the unbound state of residue conformations observed in the bound state in (a) protein-protein interactions; and (b) protein-ligand interaction. (c) The probabilities of sidechain conformations within the cores monomers. Left, center, and right panels correspond to Phe, Tyr, and Trp residues, respectively. The probabilities in the bound state (not shown) of complexes are highly correlated with those in the unbound state, implying that the monomers contribute most to conformational restriction, even in the bound state.

A likely explanation for the differences in sidechain flexibility between the designed and native structures and complexes is that evolutionary pressures on native structures involve aspects of negative design. Functioning cell circuitry, for example, requires not only that binding proteins bind their targets with high affinity, but just as importantly, that they not bind other cellular proteins, which will almost always be in vast excess over the former.¹⁵ Likewise, enzymes and ligand-binding proteins must bind the small molecules relevant to their function but not other small molecules present in the cell. From a free-energy standpoint, the preordering of sidechains in unbound complexes is beneficial as it reduces the loss of entropic degrees of freedom upon binding. The interactions in monomeric proteins must stabilize the folded state, but not the much larger number of non-native states.

How can evolutionary pressure accomplish this high selectivity for native structures and interac-

tions? To gain insight into this issue, we investigated the structural contexts of the low-probability sidechains in designed structures and complexes and the high-probability sidechains in native conformations (Supporting Information Figures). The computed energies of these low- and high-probability sidechains are all very favorable. We found that designed low-probability aromatic sidechains tend to be exposed in the unbound state and make few if any interactions. In contrast, native high-probability sidechains interact with several residues within the host monomer and are often conformationally restricted by contacts with backbone or C β atoms. Native proteins appear to have dense interaction networks between nearby sidechain and backbone atoms that favor native conformations over non-native conformations and are not found in isoenergetic designed conformations. Evolution appears to have used negative design to ensure that bulky aromatic residues are constrained to very specific

conformations within the cores of monomers and at interfaces.

Our results highlight an important aspect of evolved native structures—the conformational restriction of large sidechains to disfavor non-native interactions—and have immediate relevance to protein design. The reference set of designed structures is necessary in this case to show that conformational restriction is not a simple consequence of selection for a very low-energy structure, which would otherwise have been a quite reasonable assumption. Further comparison of the properties of native and designed complexes will undoubtedly continue to reveal the complex and intricate features engrained in native structures and complexes by evolution. On the design side, the sidechain conformational-probability metric provides a computationally tractable alternative to explicit negative design, which as noted above, becomes prohibitive when the number of alternative structures is large. The sidechains in designs can be assessed using the metric, and overly flexible aromatic sidechains can be restricted by incorporating rigidifying interactions. The metric can also be used to guide the development of design strategies which produce more conformationally restricted sidechains by construction.

Methods

Computational design of proteins that bind native proteins

Twenty-six proteins were designed to bind to surfaces on native proteins that have been cocrystallized with other proteins. The design method followed the general procedure described by Jha *et al.*¹⁶ Briefly, the coordinates of natural scaffold proteins were obtained from the Protein Data Bank and docked against the target surface using the feature-matching algorithm PatchDock.¹⁷ Subsequent iterations of RosettaDock¹⁸ and RosettaDesign² were invoked along with backbone and sidechain minimization to enhance the stability of the complex. Candidate complexes were then selected based on computed binding energy¹⁹ and surface shape complementarity²⁰ and were manually refined to increase predicted binding energy. Evaluation of sidechain-conformational probabilities was restricted to residues on the designed proteins and did not include the target proteins.

Computational design of small-molecule binding proteins

Thirty proteins were computationally redesigned to bind to the small-molecule ligands biotin and dopamine using the Rosetta enzyme-design methodology.⁶ Briefly, a set of ~250 ligand-binding scaffolds was searched for appropriate placements of amino acid binding groups for the ligand using RosettaMatch.²¹ Initial matches were refined using RosettaDesign,²

and minimization of the rest of the binding pocket. Candidate designs were chosen and characterized as described earlier.

Computational models and native structures of monomeric proteins

Sixty-one proteins of Rossmann-like folds, which were computationally *de novo* designed (unpublished), were used. The structures were built up completely from scratch, based on previously published methods.² 39 native monomeric proteins including all- β , all α , α/β , and $\alpha+\beta$ were used.

Parameters used in computing sidechain probabilities

All calculations assumed $k_B T = 0.8$. The rotamer set used in all calculations was expanded to include rotamers one standard deviation away from the base rotamers in the Dunbrack library.¹⁰

Energy function

The energy function used in all evaluations reported here was the default all-atom Rosetta energy² known as score 12.

Source code

All code was written in C++ within the Rosetta macromolecular-modeling software suite and is freely available to academic users under the Rosetta Commons agreement (<http://www.rosettacommons.org>). RosettaScripts and commandlines for running these evaluations are available as Supporting Information. Scripts for generating the histograms using gnuplot are provided as Supporting Information as well.

Acknowledgments

The authors thank Jacob E. Corn and Erik Procko for providing some of the computational designs of protein-protein complexes used in this analysis. The authors thank Paul Bates, Alexandre Bonvin, Howook Hwang, Joel Janin, Iain Moal, and Zhiping Weng for providing experimental-binding data for this analysis before publication. S.J.F. was supported by a long-term fellowship from the Human Frontier Science Program and N.K. by a Japan Society for the Promotion of Science (JSPS) Postdoctoral Fellowship for Research Abroad.

References

1. Dahiyat BI, Mayo SL (1997) De novo protein design: fully automated sequence selection. *Science* 278: 82–87.
2. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D (2003) Design of a novel globular protein fold with atomic-level accuracy. *Science* 302: 1364–1368.
3. Shifman JM, Mayo SL (2003) Exploring the origins of binding specificity through the computational redesign

- of calmodulin. *Proc Natl Acad Sci USA* 100: 13274–13279.
4. Kortemme T, Joachimiak LA, Bullock AN, Schuler AD, Stoddard BL, Baker D (2004) Computational redesign of protein-protein interaction specificity. *Nat Struct Mol Biol* 11: 371–379.
 5. Rothlisberger D, Khersonsky O, Wollacott AM, Jiang L, DeChancie J, Betker J, Gallaher JL, Althoff EA, Zanghellini A, Dym O, Albeck S, Houk KN, Tawfik DS, Baker D (2008) Kemp elimination catalysts by computational enzyme design. *Nature* 453: 190–195.
 6. Jiang L, Althoff EA, Clemente FR, Doyle L, Rothlisberger D, Zanghellini A, Gallaher JL, Betker J, Tanaka F, Barbas CF III, Hilvert D, Houk KN, Stoddard BL, Baker D (2008) De novo computational design of retro-aldol enzymes. *Science* 319: 1387–1391.
 7. Baker D (2010). An exciting but challenging road ahead for computational enzyme design. *Protein Sci* 19: 1817–1819.
 8. Watters AL, Deka P, Corrent C, Callender D, Varani G, Sosnick T, Baker D (2007) The highly cooperative folding of small naturally occurring proteins is likely the result of natural selection. *Cell* 128: 613–624.
 9. Kastriitis PL, Bonvin AM (2010) Are scoring functions in protein-protein docking ready to predict interactions? Clues from a novel binding affinity benchmark. *J Proteome Res* 9: 2216–2225.
 10. Dunbrack RL, Jr, Karplus M (1994) Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. *Nat Struct Biol* 1: 334–340.
 11. Das R, Baker D (2008) Macromolecular modeling with rosetta. *Annu Rev Biochem* 77: 363–382.
 12. Bowie JU, Luthy R, Eisenberg D (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 253: 164–170.
 13. Havranek JJ, Harbury PB (2003) Automated design of specificity in molecular recognition. *Nat Struct Biol* 10: 45–52.
 14. Richardson JS, Richardson DC, Tweedy NB, Gernert KM, Quinn TP, Hecht MH, Erickson BW, Yan Y, McClain RD, Donlan ME, Surles MC (1992) Looking at proteins: representations, folding, packing, and design. *Biophysical Society National Lecture, 1992. Biophys J* 63: 1185–1209.
 15. Zarrinpar A, Park SH, Lim WA (2003) Optimization of specificity in a cellular protein interaction network by negative selection. *Nature* 426: 676–680.
 16. Jha RK, Leaver-Fay A, Yin S, Wu Y, Butterfoss GL, Szyperski T, Dokholyan NV, Kuhlman B (2010) Computational design of a PAK1 binding protein. *J Mol Biol* 400: 257–270.
 17. Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ (2005) PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res* 33: W363–W367.
 18. Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl CA, Baker D (2003) Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J Mol Biol* 331: 281–299.
 19. Kortemme T, Kim DE, Baker D (2004) Computational alanine scanning of protein-protein interfaces. *Sci STKE* 2004:pl2.
 20. Lawrence MC, Colman PM (1993) Shape complementarity at protein/protein interfaces. *J Mol Biol* 234: 946–950.
 21. Zanghellini A, Jiang L, Wollacott AM, Cheng G, Meiler J, Althoff EA, Rothlisberger D, Baker D (2006) New algorithms and an in silico benchmark for computational enzyme design. *Protein Sci* 15: 2785–2794.