Protein Interfacial Pocket Engineering via Coupled Computational Filtering and Biological Focusing Criterion

FAISAL REZA, PEIJUN ZUO, and JINGDONG TIAN

Department of Biomedical Engineering and Institute for Genome Sciences and Policy, Duke University, 136 Hudson Hall, Box 90281, Durham, NC 27708-0281, USA

(Received 25 September 2006; accepted 11 April 2007; published online 24 April 2007)

Abstract—To engineer bio-macromolecular systems, protein–substrate interactions and their configurations need to be understood, harnessed, and utilized. Due to the inherent large numbers of combinatorial configurations and conformational complexity, methods that rely on heuristics or stochasticity, such as practical computational filtering (CF) or biological focusing (BF) criteria, when used alone rarely yield insights into these complexes or successes in (re)designing them. Here we use a coupled CF–BF criterion upon an amenable interfacial pocket (IP) of a protein scaffold complexed with its substrate to undergo residue replacement and R-group refinement (R^4) to filter out energetically unfavorable residues and R-group conformations, and focus in on those that are evolutionarily favorable. We show that this coupled filtering and focusing can efficiently provide a putative engineered IP candidate and validate it computationally and empirically. The CF–BF criterion may permit holistic understanding of the nuances of existing protein IPs and their scaffolds and facilitate bioengineering efforts to alter substrate specificity. Such approach may contribute to accelerated elucidation of engineering principles of bio-macromolecular systems.

Keywords—Protein scaffold, Interfacial pocket, DNA substrate, Complex, Residue replacement, R-group refinement, Computational bioengineering, Synthetic systems biology.

ABBREVIATIONS

BF	biological focusing
CF	computational filtering
GMEC	global minimum energy conformation
IP	interfacial pocket
PS	primary structure
QS	quaternary structure
R^4	residue replacement R-group refinement
RMSD	root mean square deviation
SS	secondary structure

TS tertiary structure
 VDW van der Waals

INTRODUCTION

Elucidating the properties of a protein interfacial pocket (IP) can be a daunting task,^{2,32} let alone re-engineering it by altering residue and R-group arrangements to endow intended new functionalities. The IP of a protein may be abstracted as a set of amino acid residues, not necessarily adjacent in the linear polypeptide sequence, that form a local biochemical environment in three-dimensional structure space using conformations of their R-groups that is favorable to binding the proper substrate.²⁹ These residues are housed amongst the rest of the residues, or scaffold, that do not participate in direct binding. Contribution to this favorable local environment can arise from a number of biophysical factors. The *steric effects*, approximated by a pairwise Lennard–Jones interaction energy, E^{VDW} , of the van der Waals (VDW) radii of atoms composing the amino acid residues of the IP as well as the proper substrate, contribute to favorable binding between them and provide hindrance and shielding against unintended side-reactions involving other substrates.⁴⁶ The *coulombic effects*, represented by an analogous pairwise electrostatic interaction energy, $E^{\text{electrostatics}}$, enable charge complementarities between regions of the IP and proper substrate and repulsive mismatches with other substrates.⁴⁶ Since both the IP and a region of the substrate with which it interfaces are occupied by electrostatic interactions with water molecules or some other solvent, evacuating this solvent is quantified as the electrostatic desolvation energy of the former, $(\Delta G^{\text{desolvation}})^{\text{IP}}$, and the latter, $\Delta G^{\text{desolvation}})^{\text{substrate}}$. Thus, an estimate has been often used for the net binding free energy, ΔG , from a linear sum of these weighted energies⁴⁶:

Address correspondence to Jingdong Tian, Department of Biomedical Engineering and Institute for Genome Sciences and Policy, Duke University, 136 Hudson Hall, Box 90281, Durham, NC 27708-0281, USA. Electronic mail: jtian@duke.edu

$$\Delta G = w_{\text{IP}}^{\text{desolvation}} \Delta G_{\text{electrostatics}}^{\text{desolvation IP}} + w_{\text{substrate}}^{\text{desolvation}} \times \Delta G_{\text{electrostatics}}^{\text{desolvation substrate}} + \sum_i w_i^{\text{VDW}} E_i^{\text{VDW}} + \sum_i w_i^{\text{electrostatics}} E_i^{\text{electrostatics}} + C$$

The protein IP engineering possibilities, as outlined in Fig. 1, and IP residue replacement R-group refinement (R^4) maintain these favorable ΔG of interactions and global minimum energy conformations (GMECs) of the protein–substrate complex as a whole.¹⁴

Computational Filtering (CF) Approaches

There are a number of CF approaches to perform R^4 with the aforementioned energetic conditions under consideration. However, due to the rapidly increasing degrees of freedom at each residue, n , of the protein chain, coupled with the specific characteristics of the 20 amino acids that can be found at each position, an colossal combinatorial quagmire of 20^n possibilities require modeling and analysis—and for an average-sized protein composed of 100 amino acids, simulating 20^{100} possible physical combinations exceeds the number of atoms in the known universe. Thus, the probability of the protein's IP locating its native state by pursuing all these combinations is biologically infeasible (known as the Levinthal paradox)¹⁵ and computationally impractical (known as the Blind Watchmaker paradox).¹⁵ An exhaustive structural

bioinformatics search for IP formation and end-state continues to be a challenge that is tackled using filtering, heuristics, homology, distributed computing, and high performance supercomputers with varied success.³⁹

Heuristics are often helpful and necessary in undertaking R^4 at the scale of IPs. For example, heuristics in genetic algorithms, mean field algorithms, constraint logic programming enumeration, or database search perform adequately under certain scenarios and assumptions and not as well with others.¹¹ While the computational cost is lessened or efficiency increased compared to the exhaustive search, the quality, however, of the end solution may or may not be consistent rather than assuring that the particular IP R^4 generated by the heuristic is located at or near the GMEC.

Homology can often aid in proper R^4 as well. Here, informatics searches and interpolations from signature sequences of a few residues composing a key motif of IP, substrate, or both can provide clues for engineering. This can extend further to domain sampling of entire regions across the protein that compose the IP. While this may be effective in well-investigated and documented systems, those sequences or structures with no similarity or availability of such information can hinder this approach. Even with fertile sources, often the R^4 is limited by what has been already observed to transpose well.^{26,27}

In similar fashion, partitioning and docking can narrow the possibilities for R^4 .²⁵ A collection of IP conformations can be generated that each present a different VDW, electrostatic profile, or desolvation cost. By docking this collection of IP conformations to the proper substrate, the affinity features of those subpartition of IPs that dock more readily can be gleaned. However, fully enumerating all the elements in this collection may be computationally difficult or biologically unsubstantiated.

Furthermore, exact filtering algorithms, among them integer programming,²¹ dead-end elimination (DEE),^{12,13,23} and A*,²⁴ can advance the R^4 process by eliminating possibilities.¹⁷ For example, in DEE the relative global energy, E_{global} , of an IP is composed of the linear sum of the energy contributions from the backbone, the self and interaction with backbone energy, $E(i_r)$, of rotamer, r , at its position, i , and its pairwise interaction energy with rotamer at nearby position, j :

$$E_{\text{global}} = E_{\text{backbone}} + \sum_i E(i_r) + \sum_i \sum_j E(i_r j_s); \quad i < j$$

Thus, if the minimum energy, determined via some given discrete rotamer library and energy function, or

FIGURE 1. Protein IP engineering possibilities. The wild-type protein, consisting of the original scaffold and IP with the original substrate (top) can serve as a starting point for three distinct engineering possibilities: an advantageous original scaffold can support an engineered IP that binds a different substrate than wild-type (bottom left); an original IP that binds the original substrate well can be adapted to an engineered scaffold (bottom center); or both IP and scaffold can be engineered to provide advantageous support and binding to a different substrate (bottom right).

best case, arrangement of rotamer, i_r , still has a higher energy than the maximum energy, or worst case, arrangement of an alternate rotamer, i_r :

$$E(i_r) + \sum_j \min_s E(i_r j_s) > E(i_r) + \sum_j \min_s E(i_j s); i \neq j$$

then the former rotamer is considered an energetic dead-end for further investigation as its arrangement is guaranteed to not be a participant in the GMEC, thus filtering the number of possibilities than need to undergo R^4 . However, this guarantee is accompanied by an expensive computational cost due to the enumeration of all elements of the rotamer library in use at each residue position of the IP. In addition, since these IP R-groups need to be energy minimized as a whole, then DEE may no longer be provably accurate. In summary, the CF approaches are often a trade-off between the quality of the end IP candidate and the efficiency to reach it.⁴²

Biological Focusing (BF) Approaches

Correspondingly, there are many BF approaches to perform R^4 so that resulting possibilities are in or near the aforementioned energetic conditions, perhaps by virtue of the constraints and fitness requirements existing in and imposed by the biological environment.^{44,45} Here, the parallel processing nature of this environment may provide a natural, even advantageous, platform to evaluate the large combinatorial number of possibilities and interdependencies to be considered in a tractable manner. However, this evaluation is often performed in a stochastic, discovery-driven investigation using various mutagenesis techniques, recombination, and directed evolution among others to screen for high performing clones or select those that survive from a large starting population representing the number of possibilities.⁷

Stochastics are often necessary for R^4 at a single position in the protein, let alone the half-dozen to a dozen residues that comprise some IPs.¹⁹ Consider a random mutagenesis methodology using mutagenic chemicals, wobble base PCR, or error prone PCR to incorporate mutations at the genetic level that will be selected or screened for the desired characteristics at the protein level. Though apparently misguided, it has been observed that non-obvious mutations can give rise to proteins with new characteristics.³

Another group of approaches to achieve R^4 using biology relies on using the recombination of existing components in the system to generate new promising possibilities.⁴³ Among these is incremental truncation to correlate the loss or gain of certain IP features and functions to the gene and protein truncation positions.^{31,38} There is also homologous gene shuffling to

generate variants of the original IP from internal wellsprings of diversity.⁹

These external and internal sources of stochastics can be considered aspects of directed and simulated evolution, which mimic the fitness requirements, survival and natural selection, propagation and amplification and of individuals, or IPs, to evaluate massive potential-filled populations with desirable properties.¹⁹ However, these stochastic approaches rely on the robustness of this evolutionary condition to propagate order from randomness. In summary, the BF approaches are usually a compromise between the intended end IP and those that arise serendipitously or survive having unintended properties.

Coupling Computational and Biological Approaches: CF-BF

While CF and BF each has its own advantages and drawbacks, a synergistic coupling of CF and BF may narrow the scope to a smaller number of high-quality, intended candidates more efficiently than either alone (Fig. 2). This smaller number of possibilities is also more amenable to downstream computational and empirical evaluation and feedback. In this research, we attempt to demonstrate the application of the CF-BF criterion to computationally engineer a putative IP on the scaffold of the restriction endonuclease R.PvuII to bind a different DNA substrate, which is to be that of the restriction endonuclease R.EcoRV.

MATERIALS AND METHODS

As the various CF approaches are described in aforementioned references, the materials and methods for the BF aspects of the CF-BF criterion are as follows. Since restriction endonucleases are a representative class of proteins with high specificity to their DNA substrates,^{33,34} the electronic .pdb files containing crystallographic coordinates of protein structures of comparative candidates were obtained via a survey of the Protein Data Bank (<http://www.rcsb.org/pdb>)⁶ and REBASE: the Restriction Enzyme Database (<http://rebase.neb.com>)³⁶ online resources. Of the 22 restriction enzymes and 10 methyltransferases with available crystal structural information, PvuII was the only candidate for which structural information in .pdb files were available for both the restriction endonuclease, R.PvuII, (denoted by the **R.** prefix) and the corresponding methyltransferase modification enzyme, M.PvuII, (denoted by the **M.** prefix), and thus were downloaded for its greater comparative specificity of the restriction-modification system for the same DNA substrate.^{1,18} Also, as seen in Table 1, given that

FIGURE 2. CF–BF reduces the search space Ω and the corresponding cost required to locate GMEC. Using existing CF criterion, shown in red arrows, the search space of all possibilities, Ω , eliminates residues and R-group configurations of those residues that are most likely not in the GMEC based on pairwise local energies, to yield a smaller number of conformational possibilities, shown in dotted blue line, that must be evaluated via global energy minimization (top panel). Coupling to a BF criterion, shown in green, improves this condition by further reducing Ω to an even smaller number of possibilities based on evolutionarily relevant residues and R-groups, to be evaluated for minimum global energy as well as functionality (bottom panel).

R.PvuII is among the smallest known restriction endonucleases available, it may be more amenable to structural bioinformatics analysis and a tolerant acceptor of IP engineering.⁸ The other candidate .pdb files downloaded were for R.EcoRI,²⁰ R.EcoRV,^{37,47} and R.BamHI³⁰ since these restriction endonucleases also had available structural information. For all these candidates, except M.PvuII, both the unbound, or native, (denoted by the N. suffix) and the bound (to DNA substrate, denoted by the D. suffix) forms were accessible, providing further comparison of conformational change with binding and a fertile source of donor IP residues for engineering. While it would have been insightful to compare isoschizomers, or enzymes that recognize the same DNA substrate and perform similarly, to R.PvuII's CAGCTG, such as R.DmaI,

there are no such available crystal structure data available at this time. Similarly lacking crystal structure data but even more useful from a comparative perspective would be neoschizomers, or enzymes that recognize the same DNA substrate but perform their activity at different positions from the prototype.³⁵

Primary structure (PS) BF analysis was performed by first querying the REBASE database for comparative protein candidates (1PVI.pdb, 1BOO.pdb, 1ERI.pdb, 4RVE.pdb, 1BHM.pdb) and extracting the source organism. For each protein and associated source organism, the identity of the corresponding oligonucleotide bases of DNA substrate and means of interaction was determined. The protein polypeptide sequences were retrieved from the Protein Data Bank for each .pdb entry and cross-checked with the SEQRES fields in the .pdb files. Furthermore, these sequences were used to construct a phylogenetic tree. Multiple sequence alignments of the sequences were generated using a pairwise alignment evolutionary distance matrix, neighbor-join clustering, and CLUSTALW algorithms.⁴⁰

Secondary structure (SS) BF analysis was performed for the remaining proteins (1PVI.pdb, 1ERI.pdb, 4RVE.pdb, 1BHM.pdb) after PS BF by querying the Protein Data Bank for “Sequence Details” section to assign SS based on the .pdb structure file’s “Author” and domain assignment using the Structural Classification of Proteins (SCOP) backend database.²⁸ The hydrophobic profiles of each protein were determined using the Kyte–Doolittle method.²²

Tertiary (TS) and quaternary (QS) structure BF analyses was performed for the remaining proteins (1PVI.pdb and 4RVE.pdb) after PS and SS BF using the open source PyMOL v.0.99 molecular graphics real-time visualization and manipulation software with embedded Python scripting and interpreter.¹⁰ Each .pdb file was loaded, preset to cartoon rendering of the polypeptide SS, TS, and QS, enabled main and side chain rendering of the oligonucleotide substrate, and directional coloring of the polypeptide chains using a spectral gamut ranging from cooler blue hues at the N-termini to warmer red hues at the C-termini. After isolating all atoms composing the hexameric recognition sequence on the sense oligonucleotide chain to serve as points of origin, the set of all residues within a 3.0 angstroms (Å) boundary from these points were selected, the average distance between hydrogen bond donor and acceptor atoms. This set was pruned of those atoms located at the origin and the antisense oligonucleotide chain, leaving the subset of atoms that were part of IP residues and R-groups within this boundary. Upon labeling, the polypeptide positions and residues at those positions were tabulated against the closest proximity oligonucleotide base. This

TABLE 1. Comparative protein candidate structures for R.PvuII.

Protein	Description	.pdb File	Structure	Size
R.PvuII-N	PvuII Native apo form restriction enzyme	1PVU		Dimer (shown) with 157 residues per monomer
R.PvuII-D	PvuII restriction enzyme with DNA substrate	1PVI		As above with DNA base pairs
M.PvuII-N	PvuII cytosine N-4 apo form Methyltransferase	1B00		Monomer (shown) with 323 residues
R.EcoRI-N	EcoRI Native apo form restriction enzyme	1QC9		Dimer with 276 residues per monomer (shown)
R.EcoRI-D	EcoRI restriction enzyme with DNA substrate	1ERI		As above with DNA base pairs
R.EcoRV-N	EcoRV Native apo form restriction enzyme	1RVE		Dimer (shown) with 245 residues per monomer
R.EcoRV-D	EcoRV restriction enzyme with DNA substrate	4RVE		As above with DNA base pairs
R.BamHI-N	BamHI Native apo form restriction enzyme	1BAM		Dimer (shown) with 213 residues per monomer
R.BamHI-D	BamHI restriction enzyme with DNA substrate	1BHM		As above with DNA base pairs

process was repeated for the antisense oligonucleotide chain with similar results, due to the palindromic nature of the DNA substrates and homodimeric nature of the restriction endonucleases. In addition to these steric VDW calculations, qualitative vacuum electrostatics assessments of the protein IP and accessible surface for each structure were generated using a local protein contact potential, without solvent dielectrics, and with equilibrium charges and radii settings from the Assisted Model Building and Energy Refinement (AMBER 99) force field to evaluate IP charge complementarity to the substrate DNA.¹⁶ Given these sterics and electrostatics, the consensus participating positions from the acceptor IP were overlaid with the

consensus participating residues from the donor IP to propose a putative engineered IP on the acceptor scaffold that binds the donor substrate.

Preliminary validation of the proposed putative engineered IP was carried out both computationally and empirically. Structural mutagenesis was performed on both monomers of R.PvuII-D using a PyMol-native rotamer library to generate the mutant homodimeric enzyme. Discrete mutant rotamers were auto-positioned based on calculated lowest energy, steric hindrance minimizing conformations, and then relaxed to assume along the same spatial direction as would be achieved by a natural, continuous R-group. Then, in one approach, GATATC DNA substrate coordinates

were extracted from R.EcoRV-D, and in the other B-form DNA substrate coordinates was generated *de novo* using NUCGEN.⁵ Each GATATC substrate was then inserted into R.PvuII-D and affine space aligned along the CAGCTG DNA substrate using the shared second A and fifth T bases as spatial and directional coordinates of reference. Upon aligning, CAGCTG DNA substrate was deleted resulting in the R.PvuII Putative Engineered IP mutant-D, i.e., the mutant complexed with GATATC DNA substrate. Hydrogen bond and polar contact patterns between the IP residues and DNA substrate involved in recognition were calculated and compared for engineered and wild-type complexes.

Preliminary empirical validation was carried out using a cell survival assay. R.PvuII Putative Engineered IP mutant was synthesized *de novo* using a similar protocol as described in⁴¹ and the correct synthetic sequence was selected and confirmed by standard DNA sequencing technology. This synthetic gene was sub-cloned into the pET-21a expression vector (Novagen), which was then transformed into *E. coli* BL21 cells. Transformed cells were cultured with appropriate selection antibiotics in the presence or absence of IPTG, which induced expression of R.PvuII Putative Engineered IP mutant.

RESULTS AND DISCUSSIONS

The coupled CF–BF criterion (Fig. 3), can be used to generate a few promising candidate engineered IPs in the original scaffold that will bind to a different substrate. Upon freezing those scaffold coordinates not relevant to engineering, the CF can be calibrated to both the donor and acceptor IPs so that it can appropriately eliminate energetically impossible or unlikely residue participants, to produce some semi-engineered IP candidates that may have the desired binding ability. But given the size of many IPs, there will still be too many to fully consider energy minimization. This number can be reduced further by BF on those aspects of the candidate and other IPs shown to be evolutionarily conserved or fit. Note that this BF permits informed selection of R^4 from a continuous set of remaining possible R-group conformations rather than discrete enumeration from a rotamer library often used in the CF performed upstream. Also note that this BF is not conditional upon CF and can be used standalone or coupled to the latter to yield results. This smaller number of engineered IPs will still be subject to downstream energy minimization as is the rest of the protein–DNA complex. However, given their reduced number, a larger portion or all of them can now be evaluated.

FIGURE 3. Flowchart of coupled CF–BF criterion to engineer an IP. This flow of the CF–BF criterion demonstrates the three possibilities from Fig. 1 to engineer an IP on the original scaffold that binds to a different substrate than that of the original protein. Should the CF and BF determine that the two proteins behave structurally or interfacially similar, it may be possible for one to act as a IP donor, having key residues and R-groups conformations of those residues, that can be transplanted into the other protein, the IP acceptor.

Primary Structure (PS) BF

The objective of BF on PS is to focus on those candidates that have similar origins based on their polypeptide sequences. Candidate proteins are all from prokaryotic organisms (Table 2). Yet, it is notable that all but M.PvuII-N perform a similar biological function—sequence recognition and cleavage of DNA substrates. While a multiple sequence alignment of polypeptide sequences is not particularly revealing (Fig. 4), the associated phylogenetic tree indicates that the methyltransferase M.PvuII-N (1B00.pdb) which protects DNA from cleavage is most distant from the others. While this tree does not represent actual evolutionary patterns, it is not surprising that the monomeric M.PvuII-N may have arisen differently than the dimeric restriction endonucleases. Thus, this PS BF deemphasizes the IP of this methyltransferase as a source of donor residues and R-group conformations for R⁴ into the R.PvuII-N/D dimer.

Secondary Structure (SS) and Hydropathy BF

The objective of BF on SS is to focus on those remaining candidates that have similar local hydrophobic profiles and conformations of the IP polypeptide backbone, such as the alpha helix and beta sheet. In addition, given evolutionary fold conservation at protein IPs, such as active and allosteric sites, it may be worthwhile to compare how these local conformations interact with the DNA substrate. Also, this conservation may influence the choices made in R⁴ since certain residues are more capable in participating in particular SS, as they are able to adopt the necessary backbone dihedral angles. A mapping of these dihedral angles to the corresponding SS and capable residues can be found in a Ramachandran plot.

For this analysis, hydropathic profiles remained uninformative, but similarity in SS motif interactions

TABLE 2. Origins and attributes of chosen proteins.

Protein	Source organism	DNA substrate
R.PvuII-N	<i>Proteus vulgaris</i> (ATCC 13315)	5'-CAG ↓ CTG-3' 3'-GTC ↑ GAC-5'
R.PvuII-D	As above	As above
M.PvuII	As above	5'-CAG ^{m4} CTG-3' 3'-GTC _{m4} GAC-5'
R.EcoRI-N	<i>E. coli</i> RY13 (R.N. Yoshimori)	5'-G ↓ AATTC-3' 3'-CTTAA ↑ G-5'
R.EcoRI-D	As above	As above
R.EcoRV-N	<i>E. coli</i> J62 pLG74 (L.I. Glatman)	5'-GAT ↓ ATC-3' 3'-CTA ↑ TAG -5'
R.EcoRV-D	As above	As above
R.BamHI-N	<i>Bacillus amyloliquefaciens</i> H (ATCC 49763)	5'-G ↓ GATCC-3' 3'-CCTAG ↑ G-5'
R.BamHI-D	As above	As above

FIGURE 4. PS properties of chosen proteins. The PSs of the chosen proteins were compared using multiple sequence alignment (top panel) and associated phylogenetic tree (bottom panel).

to DNA grooves permitted further focusing (Supplementary Fig. 1). It was calculated that both R.EcoRI-N/D and R.BamHI-N/D tend to have a greater proportion and longer stretches of alpha helices (shown as waves) than beta sheets (shown as arrows), while both R.EcoRV-N/D and R.PvuII-N/D are more balanced in beta sheet content. Notably, both R.EcoRI-N/D and R.BamHI-N/D approach and recognize the DNA substrate from the major groove via an alpha helix and a loop and produce 5' sticky ends, while both R.EcoRV-N/D and R.PvuII-N/D do so from the minor groove via a beta sheet and beta-like turn and produce blunt or 3' sticky ends. Thus, this SS BF deemphasizes R.EcoRI-N/D and R.BamHI-N/D as sources for IP donation to R.PvuII-N/D, leaving R.EcoRV-N/D as a more biologically promising candidate.

Tertiary (TS) and Quaternary (QS) Structure BF

The objective of BF on TS and QS is to spatially align the remaining two well-focused candidates and

readily identify the IP residues close enough to interact with the substrate. Given that both R.PvuII-N and R.EcoRV-N recognize and bind to a uniform, helical substrate, this can facilitate R⁴ by acting as a common coordinate reference from which residues from the donor IP can be mapped onto positions on the acceptor IP. A promising mapping can be confirmed using qualitative vacuum electrostatics assessments of the protein IP and accessible surface, since the engineered IP should have the electrostatic profile of the donor IP while the rest of the scaffold should remain as is.

Longitudinal and axial views of the DNA substrate were presented against each protein for orientation purposes (upper panels of Fig. 5). From the longitudinal views, some asymmetric distortion of the DNA substrate upon binding can be seen. This may contribute to the lack of identical IP residues found within the 3.0 Å boundaries for both the sense and antisense strands of the DNA substrate. Upon selecting a DNA strand as points of origin, the IP residues within this 3.0 Å boundary are closest to which DNA base are explicitly indicated (lower panels of Fig. 5). Taking advantage of the existing symmetries, the consensus IP position 140 on R.PvuII-D was determined to interact with the first base of the DNA substrate, and so on. Similar symmetries showed a consensus IP residue LYS on R.EcoRV-D to interact with the first base of the DNA substrate. By integrating discrete levels of biological abstraction, from PS to SS to TS and to QS,

in a systematic manner, a putative engineered IP is mapped on the original R.PvuII-D scaffold (Fig. 6) to confer specificity and bind the R.EcoRV GATATC substrate.

Computational Validation via Structural Mutagenesis

The putative engineered IP mapping was validated structurally using relaxed rotamer-based mutagenesis and hydrogen bond and polar contact pattern comparison. Discrete mutant rotamers were positioned with greater than majority occupancy at that lowest energy with little to no steric hindrance. After positioning, local relaxing permitted the mutant to assume similar spatial direction as wild-type residue had, thus achieving continuous R-group positioning. Affine space alignment repositioned the GATATC DNA substrate from the R.EcoRV-D orthogonal coordinate system to that of the R.PvuII Putative Engineered IP

R.PvuII Putative Engineered IP mutant-D		
residues within 3.0 Å of substrate DNA		
	sense	anti-sense
G	LYS-140	LYS-140
A	ASN-84	ASN-84
T	THR-93	THR-93
A	LYS-33, THR-35	LYS-33, THR-35
T	THR-83	THR-83
C	ASN-141, THR-143	ASN-141, THR-143

having consensus residues of R.EcoRV to bind GATATC substrate, at consensus 3.0 Å positions of R.PvuII

FIGURE 6. CF-BF for putative engineered IP on original R.PvuII-D scaffold to bind EcoRV-D GATATC substrate.

FIGURE 5. TS and QS properties of remaining proteins after primary and SS focusing.

mutant-D complex and then aligned it to the CAGCTG DNA substrate. Though the R.EcoRV-D GATATC DNA substrate was slightly more bent, and the B-form one slightly less, than the R.PvuII-D CAGCTG DNA substrate, the affine space alignments achieved good fits with expected deviations occurring at the substrate extremes, with root mean square deviations (RMSD) of 2.02 and 2.07 Å for the R.EcoRV-D and canonical B-form conformations, respectively. Hydrogen bonding and polar contacts were made by the mutant IP residues to all bases and backbone in both conformations of GATATC substrate, just as for the wild-type complexes (Fig. 7). Thus, this suggests that the mutant is labile yet specific enough to approach and make recognition contacts with B-form DNA, and then maintain these contacts while bending it like R.EcoRV-D to expose the scissile phosphate for enzymatic cleavage. Computationally, this R.PvuII Putative Engineered IP mutant exhibited

promising recognition characteristics and was further evaluated empirically.

Preliminary Empirical Validation via Cell Survival Assay

A cell survival assay was performed to determine whether the engineered R.PvuII has any enzymatic activity in cutting DNA. A synthetic R.PvuII Putative Engineered IP mutant was cloned into pET-21a expression vector and cultured on agar plate in the presence or absence of IPTG. Without IPTG, the cells grew normally and formed colonies; with IPTG, cells did not grow and no colonies were found on the plate (data not shown). The result of this cell survival assay suggested that the expression of the R.PvuII Putative Engineered IP mutant in *E. coli* lead to cell death presumably due to digestion of the host chromosomal DNA.⁴ Further characterization of the mutant enzyme

FIGURE 7. Computational validation based on hydrogen bond and polar contacts with respect to steric hindrance patterns. Wild-type R.PvuII-D interacts with DNA substrate CAGCTG (top left), R.EcoRV-D with GATATC (top right), and R.PvuII Putative Engineered IP mutant also with canonical B-DNA-like (bottom left) and R.EcoRV-D-like (bottom right) GATATC in order to make hydrogen bond and polar contacts while avoiding steric hindrance. DNA are color coded according to bases (alanine = aquamarine, cytosine = crimson red, guanine = green, thymine = tan), IP residues according to standard Corey, Pauling, Kuffin (CPK) atom colors (carbon = white, oxygen = red, nitrogen = blue), and hydrogen bond and polar contacts highlighted in dotted lines (contacts from IP residues to any other atoms in yellow, from DNA hexamer to any other atoms in orange).

is underway to determine its specificity activity and substrate specificity.

CONCLUSIONS

While CF–BF produced a promising putative engineered IP, the computational and empirical validations confirmed its properties to a certain degree. The computational validation of R.PvuII-D scaffold with the engineered IP examined the hydrogen bonding and polar contact patterns to show they, like the R.EcoRV-D IP, interact with all the bases in the substrate GA-TATC DNA. Preliminary empirical validation using the cell survival assay suggested that the R.PvuII Putative Engineered IP has enzymatic activity in cutting DNA. The engineered specificity remains to be determined by further biochemical assays. These validations, in turn, will inform and improve the CF–BF criterion through further iterations of this process for these types of nucleic acid binding enzymes.

The IP engineering possibilities are nearly as vast as the diversity of biology itself. Using CF alone, R^4 of a hexameric DNA recognition site and a monomer of a dimer restriction enzyme, assuming that there is one residue interacting with each base of DNA and a limited number of rotamers in a library represents all possible R-group conformations of the 20 naturally occurring residues, would have required the evaluation of possible IP sequences equal to the size of the rotamer library to the sixth power in number. The CF–BF reduces this to a subset of the 20 naturally occurring residues (the subset being the 10 polar and charged residues in herein restriction endonuclease experiment) and R-group conformations, which are known to participate in the IP and interact with a similarly structured substrate. The benefits of successful IP engineering are equally numerous. Engineered IPs may lead to programmable proteins, such as restriction endonucleases that not only act as research tools, by enabling targeting, mapping and manipulation of genes and genomes, but as clinical technologies as well, by facilitating cleaving out a disease gene and repairing it with a working version *in vivo*.

ELECTRONIC SUPPLEMENTARY MATERIAL

The online version of this article (doi: [10.1007/s10439-007-9316-8](https://doi.org/10.1007/s10439-007-9316-8)) contains supplementary material, which is available to authorized users.

ACKNOWLEDGMENTS

This work was supported by a Biotechnology Pre-doctoral Training Fellowship to F.R. from NIH Grant

GM08555 and partially by the Arnold and Mabel Beckman Foundation Young Investigator Award to J.T. Further support to F.R. was provided by the Duke University Institute for Genome Sciences and Policy, Computational Biology and Bioinformatics Program.

REFERENCES

- Adams, G. M., and R. M. Blumenthal. The PvuII DNA (cytosine-N4)-methyltransferase comprises two trypsin-defined domains, each of which binds a molecule of S-adenosyl-L-methionine. *Biochemistry* 36:8284–8292, 1997.
- Akutsu, T. editor. NP-hardness results for protein side-chain packing. *Genome Informatics* 8:180–186, 1997.
- Arnold, F. H., and A. A. Volkov. Directed evolution of biocatalysts. *Curr. Opin. Chem. Biol.* 3:54–59, 1999.
- Balendiran, K., J. Bonventre, R. Knott, W. Jack, J. Benner, I. Schildkraut, and J. E. Anderson. Expression, purification, and crystallization of restriction endonuclease PvuII with DNA containing its recognition site. *Proteins* 19:77–79, 1994.
- Bansal, M., D. Bhattacharyya, and B. Ravi. NUPARM and NUCGEN: software for analysis and generation of sequence dependent nucleic acid structures. *Comput. Appl. Biosci.* 11:281–287, 1995.
- Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Res.* 28:235–242, 2000.
- Brakmann, S. Discovery of superior enzymes by directed molecular evolution. *ChemBiochem* 2:865–871, 2001.
- Cheng, X., K. Balendiran, I. Schildkraut, and J. E. Anderson. Structure of PvuII endonuclease with cognate DNA. *Embo. J.* 13:3927–3935, 1994.
- Cochran, J. R., Y. S. Kim, S. M. Lippow, B. Rao, and K. D. Wittrup. Improved mutants from directed evolution are biased to orthologous substitutions. *Protein Eng. Des. Sel.* 19:245–253, 2006.
- DeLano, W. L. The PyMOL Molecular Graphics System. DeLano Scientific, Palo Alto, CA, USA. February 6, 2006, <http://www.pymol.org>.
- Desjarlais, J. R., and N. D. Clarke. Computer search algorithms in protein modification and design. *Curr. Opin. Struct. Biol.* 8:471–475, 1998.
- Desmet, J., M. De Maeyer, and I. Lasters. Theoretical and algorithmical optimization of the dead-end elimination theorem. *Pac. Symp. Biocomput.* 2:122–133, 1997.
- Desmet, J., M. D. Maeyer, B. Hazes, and I. Lasters. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* 356:539–542, 1992.
- Dill, K. A. Dominant forces in protein folding. *Biochemistry* 29:7133–7155, 1990.
- Dill, K. A. Polymer principles and protein folding. *Protein Sci.* 8:1166–1180, 1999.
- Duan, Y., C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, and T. Lee, *et al.* A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* 24:1999–2012, 2003.
- Georgiev, I., R. H. Lilien, and B. R. Donald. Improved pruning algorithms and divide-and-conquer strategies for

- dead-end elimination, with application to protein design. *Bioinformatics* 22:e174–e183, 2006.
- ¹⁸Gong, W., M. O'Gara, R. M. Blumenthal, and X. Cheng. Structure of pvu II DNA-(cytosine N4) methyltransferase, an example of domain permutation and protein fold assignment. *Nucleic Acids Res.* 25:2702–2715, 1997.
 - ¹⁹Hellings, H. W., and F. M. Richards. Optimal sequence selection in proteins of known structure by simulated evolution. *Proc. Natl. Acad. Sci. USA* 91:5803–5807, 1994.
 - ²⁰Kim, Y. C., J. C. Grable, R. Love, P. J. Greene, and J. M. Rosenberg. Refinement of Eco RI endonuclease crystal structure: a revised protein chain tracing. *Science* 249:1307–1309, 1990.
 - ²¹Kingsford, C. L., B. Chazelle, and M. Singh. Solving and analyzing side-chain positioning problems using linear and integer programming. *Bioinformatics* 21:1028–1036, 2005.
 - ²²Kyte, J., and R. F. Doolittle. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 157:105–132, 1982.
 - ²³Lasters, I., M. De Maeyer, and J. Desmet. Enhanced dead-end elimination in the search for the global minimum energy conformation of a collection of protein side chains. *Protein Eng.* 8:815–822, 1995.
 - ²⁴Leach, A. R., and A. P. Lemon. Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. *Proteins* 33:227–239, 1998.
 - ²⁵Lilien, R. H., B. W. Stevens, A. C. Anderson, and B. R. Donald. A novel ensemble-based scoring and search algorithm for protein redesign and its application to modify the substrate specificity of the gramicidin synthetase a phenylalanine adenylation enzyme. *J. Comput. Biol.* 12:740–761, 2005.
 - ²⁶Looger, L. L., and H. W. Hellinga. Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: implications for protein design and structural genomics. *J. Mol. Biol.* 307:429–445, 2001.
 - ²⁷Luscombe, N. M., and J. M. Thornton. Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *J. Mol. Biol.* 320:991–1009, 2002.
 - ²⁸Murzin, A. G., S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247:536–540, 1995.
 - ²⁹Najmanovich, R., J. Kuttner, V. Sobolev, and M. Edelman. Side-chain flexibility in proteins upon ligand binding. *Proteins* 39:261–268, 2000.
 - ³⁰Newman, M., T. Strzelecka, L. F. Dorner, I. Schildkraut, and A. K. Aggarwal. Structure of restriction endonuclease bamHI phased at 1.95 Å resolution by MAD analysis. *Structure* 2:439–452, 1994.
 - ³¹Ostermeier, M., A. E. Nixon, and S. J. Benkovic. Incremental truncation as a strategy in the engineering of novel biocatalysts. *Bioorg. Med. Chem.* 7:2139–2144, 1999.
 - ³²Pierce, N. A., and E. Winfree. Protein design is NP-hard. *Protein Eng.* 15:779–782, 2002.
 - ³³Pingoud, A. *Restriction Endonucleases*. Berlin; New York: Springer, 2004.
 - ³⁴Pingoud, A., M. Fuxreiter, V. Pingoud, and W. Wende. Type II restriction endonucleases: structure and mechanism. *Cell. Mol. Life Sci.* 62:685–707, 2005.
 - ³⁵Roberts, R. J. Restriction enzymes and their isoschizomers. *Nucleic Acids Res.* 17(Suppl):r347–r387, 1989.
 - ³⁶Roberts, R. J., T. Vincze, J. Posfai, and D. Macelis. REBASE—restriction enzymes and DNA methyltransferases. *Nucleic Acids Res.* 33:D230–D232, 2005.
 - ³⁷Schottler, S., C. Wenz, T. Lanio, A. Jeltsch, and A. Pingoud. Protein engineering of the restriction endonuclease EcoRV—structure-guided design of enzyme variants that recognize the base pairs flanking the recognition site. *Eur. J. Biochem.* 258:184–191, 1998.
 - ³⁸Song, J. K., B. Chung, Y. H. Oh, and J. S. Rhee. Construction of DNA-shuffled and incrementally truncated libraries by a mutagenic and unidirectional reassembly method: changing from a substrate specificity of phospholipase to that of lipase. *Appl. Environ. Microbiol.* 68:6146–6151, 2002.
 - ³⁹Tanimura, R., A. Kidera, and H. Nakamura. Determinants of protein side-chain packing. *Protein Sci.* 3:2358–2365, 1994.
 - ⁴⁰Thompson, J. D., D. G. Higgins, and T. J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22:4673–4680, 1994.
 - ⁴¹Tian, J., H. Gong, N. Sheng, X. Zhou, E. Gulari, X. Gao, and G. Church. Accurate multiplex gene synthesis from programmable DNA microchips. *Nature* 432:1050–1054, 2004.
 - ⁴²Voigt, C. A., D. B. Gordon, and S. L. Mayo. Trading accuracy for speed: a quantitative comparison of search algorithms in protein sequence design. *J. Mol. Biol.* 299:789–803, 2000.
 - ⁴³Voigt, C. A., C. Martinez, Z. G. Wang, S. L. Mayo, and F. H. Arnold. Protein building blocks preserved by recombination. *Nat. Struct. Biol.* 9:553–558, 2002.
 - ⁴⁴Voigt, C. A., S. L. Mayo, F. H. Arnold, and Z. G. Wang. Computational method to reduce the search space for directed protein evolution. *Proc. Natl. Acad. Sci. USA* 98:3778–3783, 2001.
 - ⁴⁵Voigt, C. A., S. L. Mayo, F. H. Arnold, and Z. G. Wang. Computationally focusing the directed evolution of proteins. *J. Cell. Biochem. (Suppl 37)*:58–63, 2001.
 - ⁴⁶Wang, T., S. Tomic, R. R. Gabdoulline, and R. C. Wade. How optimal are the binding energetics of barnase and barstar? *Biophys. J.* 87:1618–1630, 2004.
 - ⁴⁷Winkler, F. K., D. W. Banner, C. Oefner, D. Tsernoglou, R. S. Brown, S. P. Heathman, R. K. Bryan, P. D. Martin, K. Petratos, and K. S. Wilson. The crystal structure of EcoRV endonuclease and of its complexes with cognate and non-cognate DNA fragments. *Embo J.* 12:1781–1795, 1993.