OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Computational Design of a PDZ Domain Peptide Inhibitor that Rescues CFTR Activity

Kyle E. Roberts[1], Patrick R. Cushing[2], Prisca Boisguerin[3], Dean R. Madden[2], Bruce R. Donald[1,4]*

1 Department of Computer Science, Duke University, Durham, North Carolina, United States of America, **2** Department of Biochemistry, Dartmouth Medical School, Hanover, New Hampshire, United States of America, **3** Institute for Medical Immunology, Charite Universitätsmedizin, Berlin, Germany, **4** Department of Biochemistry, Duke University Medical Center, Durham, North Carolina, United States of America

Abstract

The cystic fibrosis transmembrane conductance regulator (CFTR) is an epithelial chloride channel mutated in patients with cystic fibrosis (CF). The most prevalent CFTR mutation, ΔF508, blocks folding in the endoplasmic reticulum. Recent work has shown that some ΔF508-CFTR channel activity can be recovered by pharmaceutical modulators (''potentiators'' and ''correctors''), but ΔF508-CFTR can still be rapidly degraded via a lysosomal pathway involving the CFTR-associated ligand (CAL), which binds CFTR via a PDZ interaction domain. We present a study that goes from theory, to new structure-based computational design algorithms, to computational predictions, to biochemical testing and ultimately to epithelial-cell validation of novel, effective CAL PDZ inhibitors (called ''stabilizers'') that rescue ΔF508-CFTR activity. To design the ''stabilizers'', we extended our structural ensemble-based computational protein redesign algorithm K^* to encompass protein-protein and protein-peptide interactions. The computational predictions achieved high accuracy: all of the top-predicted peptide inhibitors bound well to CAL. Furthermore, when compared to state-of-the-art CAL inhibitors, our design methodology achieved higher affinity and increased binding efficiency. The designed inhibitor with the highest affinity for CAL (kCAL01) binds six-fold more tightly than the previous best hexamer (iCAL35), and 170-fold more tightly than the CFTR C-terminus. We show that kCAL01 has physiological activity and can rescue chloride efflux in CF patient-derived airway epithelial cells. Since stabilizers address a different cellular CF defect from potentiators and correctors, our inhibitors provide an additional therapeutic pathway that can be used in conjunction with current methods.

Citation: Roberts KE, Cushing PR, Boisguerin P, Madden DR, Donald BR (2012) Computational Design of a PDZ Domain Peptide Inhibitor that Rescues CFTR Activity. PLoS Comput Biol 8(4): e1002477. doi:10.1371/journal.pcbi.1002477

Editor: Giorgio Colombo, Consiglio Nazionale delle Ricerche, Italy

Received October 15, 2011; **Accepted** February 27, 2012; **Published** April 19, 2012

Funding: This work was supported in part by grants from the National Institutes of Health (R01 GM-78031 to B.R.D. and R01-DK075309 to D.R.M.), from the Hitchcock Foundation (to D.R.M), from the Deutsche Forschungsgemeinschaft (VO 885/3 2 to P.B.), and from the German Cystic Fibrosis Foundation Mukoviszidose e.V. (S05/08 to P.B.). Additional support was provided by NIH grants P20-GM103413 and P20-RR018787. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: brd+plos11@cs.duke.edu

Introduction

Protein-peptide interactions (PPIs) are vital for cell signaling, protein trafficking and localization, gene expression, and many other biological functions. The PDZ (PSD-95, discs large, zonula occludens-1) family of proteins forms PPIs that play crucial physiological roles, including synapse formation [1] and epithelial cell polarity and proliferation [2]. The common PDZ structural core generally binds a specific sequence motif at the extreme C-terminus of its binding partner through β-sheet interactions (Fig. 1A). Recently, key PPIs have been discovered linking the trafficking of the cystic fibrosis transmembrane conductance regulator (CFTR) to PDZ domain containing proteins [3] (Fig. 1B). Specifically, the PDZ domain of the CFTR-associated ligand (CAL) binds CFTR, targeting it for lysosomal degradation and reducing its half-life at the plasma membrane [4,5].

CFTR is an epithelial chloride channel that is mutated in cystic fibrosis (CF) patients. The most common disease-associated mutation, ΔF508-CFTR, is a single amino acid deletion that causes CFTR misfolding and endoplasmic reticulum-associated (ER) degradation. There is now evidence that the ΔF508-CFTR loss of function can be pharmacologically improved through the use of ''correctors'' [6] and ''potentiators'' [7]. Correctors, such as corr-4a [6,8], work by correcting the folding defect of CFTR and preventing ER retention of CFTR. Potentiators combat mutant CFTR gating defects and increase the flow of ions through CFTR channels present at the cellular membrane. Despite these interventions, the half-life of ΔF508-CFTR in the membrane is still reduced compared to that of the wild-type protein [9]. However, the CAL-mediated degradation of ΔF508-CFTR can be reduced by RNA interference or by mutagenesis of the CAL PDZ domain, suggesting that a competitive inhibitor of the CAL binding site could act as a CFTR ''stabilizer'' and thus ameliorate CF symptoms [3,10]. Since stabilizers address a different underlying CF defect than correctors and potentiators, combined application can achieve additive rescue of ΔF508-CFTR activity [11].

Since PDZ domains have an inherent affinity for peptides, here we focus on the use of protein design methods to rationally design a competitive peptide inhibitor that could serve as a ΔF508-CFTR stabilizer. Indeed, the development of successful peptide inhibitor design tools would provide a means to target a wide variety of PPIs for both mechanistic and therapeutic applications. Several aspects of our new K^* design algorithm (described below) are well suited to the requirements of this class of problems.

Author Summary

Cystic fibrosis (CF) is an inherited disease that causes the body to produce thick mucus that clogs the lungs and obstructs the breakdown and absorption of food. The cystic fibrosis transmembrane conductance regulator (CFTR) is mutated in CF patients, and the most common mutation causes three defects in CFTR: misfolding, decreased function, and rapid degradation. Drugs are currently being studied to correct the first two CFTR defects, but the problem of rapid degradation remains. Recently, key protein-protein interactions have been discovered that implicate the protein CAL in CFTR degradation. Here we have developed new computational protein design algorithms and used them to successfully predict peptide inhibitors of the CAL-CFTR interface. Our algorithm uses a structural ensemble-based evaluation of protein sequences and conformations to calculate accurate predictions of protein-peptide binding affinities. The algorithm is general and can be applied to a wide variety of protein-protein interface designs. All of our designed inhibitors bound CAL with high affinity. We tested our top binding peptide and observed that the inhibitor could successfully rescue CFTR function in CF patient-derived epithelial cells. Our designed inhibitors provide a novel therapeutic path which could be used in combination with existing CF therapeutics for additive benefit.

In general, structure-based computational protein design seeks amino-acid sequences that are compatible with a specific protein fold. Often, additional functional constraints are applied to the problem in order to design a protein with a given binding or catalytic activity. Because protein conformational space is large, design algorithms often assume a fixed backbone conformation and reduce side-chain configuration space by using discrete conformations called *rotamers* [12–15]. Thus, most current design methods try to solve the traditional design problem, which can be defined as: for a given *input model* (protein structure, rotamer library, and energy function), find the side chain rotamers that yield a single, global minimum energy conformation (GMEC) for the entire protein [16–34]. However, in reality, a protein in solution exists as a thermodynamic ensemble and not just a single low-energy structure [35]. Accounting for such ensembles can help find true native protein structures [36–39]. The design algorithm we present here, K^*, takes this into account by computing Boltzmann-weighted partition functions over structural molecular ensembles to find provably-accurate approximations to the binding constant for a protein complex [40,41]. The value of this approach is reflected in previous applications of the K^* algorithm to design a switch in enzyme specificity for an enzyme in the non-ribosomal peptide synthetase pathway [40] and to predict resistance mutations for antibiotic targets [42].

As with the established K^* algorithm, most successful protein design studies have focused on protein/small molecule systems, since predicting PPI binding is more challenging than small molecule binding, due to PPIs' much larger, flexible, and energetically shallow binding surfaces. The methodologies that have been developed to study protein-protein interactions and, more specifically, PDZ domain interactions, can be divided into sequence- [43,44] and structure-based [38,45–49] methods. Sequence-based methods require a large amount of sequence and binding information for the protein family and do not provide direct structural information on the modeled interaction. Among the previous structure-based alternatives, most focus on finding the single GMEC conformation, although one study suggests that designing to a set of different backbone conformations can improve recovery of PDZ domain binding motifs [45]. In addition, only the work of Altman *et al.* [46] utilizes provable techniques, and none use both provable techniques and protein ensembles. In comparison, the K^* algorithm is more general, requiring only a starting template structure and preserving structural information on the modeled interaction. It also evaluates energy-weighted ensembles, employs provable guarantees for finding the optimal sequence, and uses the minimization aware dead-end elimination (minDEE) pruning criteria [16,41] to permit continuous minimization of rotamers during the search. As a result, K^* complements existing approaches while addressing some of their methodological limitations. Here we report the development of new extensions to the K^* algorithm, enabling the software to design novel PPIs.

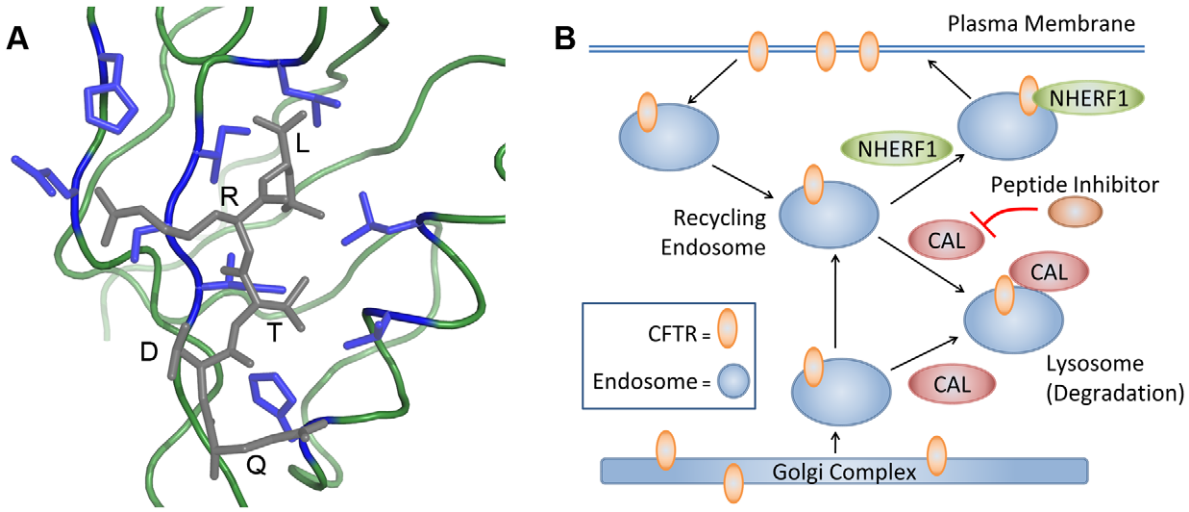

Figure 1. (A) Structural model of the CAL PDZ domain (green and blue) bound to a CFTR C-terminus mimic (gray) used as input for computational designs (PDB id: 2LOB). Residues shown in blue were modeled as flexible during the design search. (B) Model of the CFTR trafficking pathway with PDZ domain containing proteins NHERF1 and CAL. CAL is associated with lysosomal degradation of CFTR, while NHERF1 is associated with insertion of CFTR into the cell membrane.
doi:10.1371/journal.pcbi.1002477.g001

Using this new tool we designed high-affinity CAL PDZ inhibitors and validated them in both biochemical and cell-culture experiments. We present peptide array data which shows that CAL binds a specific sequence motif, but does not bind all sequences within that motif. Therefore, it is important that the K^* algorithm is able to differentiate the affinities of peptides that share the motif, rather than just separating motif from non-motif sequences. Overall, K^* searched 2166 peptide inhibitor sequences within the CAL binding motif (approximately 10^{15} possible conformations) and generated top-ranked peptides that had up to a 170-fold improvement in binding to CAL compared to the wild-type CFTR sequence. The best binder was able to rescue ΔF508-CFTR function in human cells.

Materials and Methods

K^* Algorithm

K^* computationally searches over peptide amino acid substitutions (mutations) for a given protein-peptide complex and assigns each candidate sequence a score, called a K^* *score* [40,41]. To compute the score for a given protein-peptide complex candidate sequence, K^* evaluates the low-energy conformations for the sequence and uses them to compute a Boltzmann-weighted partition function. Partition functions are computed for each protein binding partner using rotamer-based ensembles defined as $q_A = \sum_{a \in A} \exp(-E_a/RT)$, $q_B = \sum_{b \in B} \exp(-E_b/RT)$, $q_{AB} = \sum_{ab \in AB} \exp(-E_{ab}/RT)$ where q_{AB} is the partition function for protein A bound to protein B, and q_A and q_B are the partition functions for the unbound proteins, A and B. The K^* score is defined as the ratio of partition functions: $K^* = \frac{q_{AB}}{q_A q_B}$, which is an approximation of the protein complex association constant, K_A [41]. Candidate sequences are ranked based on their K^* score, where sequences with a higher K^* score are considered to have a higher affinity for the target protein.

The K^* algorithm has been described previously [16,40,41]. Briefly, to calculate a partition function for a given sequence, K^* finds low energy conformations by performing a rotamer search as follows. First, K^* uses an enhanced version of dead-end elimination (DEE), minDEE [16,41,50], to prune side-chain rotamers that provably cannot be part of low-energy structures. Since rigid-rotamer DEE [34,51] often eliminates rotamers and sequences that are involved in *bona fide* low-energy conformations [50], K^* prunes rotamers using minDEE, which allows local side-chain rotamer minimization to relieve clashes that are incorrectly pruned by rigid rotamer design methods. In order for minDEE to account for minimization during the rotamer search, it computes energy lower bounds for each rotamer pair. The branch-and-bound algorithm A^* [30] is used to enumerate conformations in gap-free order of their minimum energy bounds. These conformations are minimized and their Boltzmann-weighted energy is incorporated into the partition function. The partition function is computed with respect to the input model (protein structure, energy function, and rotamer library), so the accuracy of the partition function is bounded by the accuracy of the input model. Refer to Fig. 2 to see the general framework for the K^* algorithm.

The energy minimization scheme that is used for both the energy lower bounds computation and the minimization of a full conformation is similar to previous descriptions [41]. The K^* algorithm's minimization protocol separates a protein's degrees of freedom (DOF) into three categories: (1) backbone dihedrals (ϕ and ψ angles) (2) side-chain dihedrals (up to four χ angles per side chain) and (3) rigid body rotation and translation ($\mathbb{R}^3 \times \mathrm{SO}(3)$). The minimization process holds the backbone dihedrals fixed

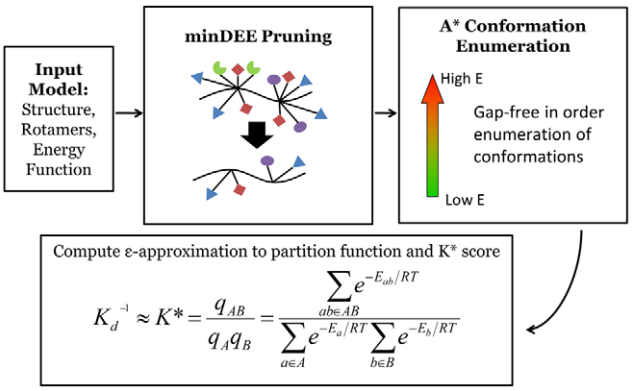

Figure 2. Overview of K^* Algorithm. The K^* algorithm searches over protein sequences and conformations to find the protein complexes with the best binding constant. K^* takes an *input model* composed of an initial protein structure, a rotamer library to search over side-chain conformations, and an energy function to evaluate conformations. Minimization-aware DEE (minDEE) prunes rotamers that are not part of the lowest energy conformations for a given sequence. The remaining conformations from minDEE are enumerated in order of increasing energy lower bounds using A*. Finally, the conformations are Boltzmann-weighted and used to compute partition functions and ultimately a K^* score for each sequence.
doi:10.1371/journal.pcbi.1002477.g002

while allowing the side-chain dihedral and rigid body DOF to minimize. The minimization over these DOF is performed using gradient descent. To prevent rotamers from minimizing from one rotamer to another, each side-chain dihedral was only allowed to move a maximum of 9° from its modal rotameric value.

Extension of K^* to Amino Acid Substitutions/Flexibility on Two Protein Strands

K^* relies on the mathematically provable guarantees of each of its steps (Fig. 2) to compute an accurate K^* score. If we were to use heuristic steps to find the low energy conformations, it could not be guaranteed that all the low energy conformations are found and we would lose the ability to calculate a provably-good ε-approximation (where ε is user-defined) to each partition function for the design system. Because of the provable aspects of K^*, if K^* makes an errant prediction, we can be certain that it is due to an inaccuracy in the input model and not a problem (such as inadequate optimization) with our search algorithm. This makes it substantially easier to improve the model based on experimental feedback, as we show in Section S2 of Text S1.

Before applying K^* to PPI designs, we first had to ensure that the mathematical framework of K^* could be extended to cover larger systems. For large designs such as PPIs, the provable guarantees of K^* no longer hold as they did for small design systems. Specifically, the previous K^* proofs [41] for intermutation pruning and guaranteeing the accuracy of the K^* score, relied on properties of small molecule design systems that are not true for PPIs. We now show that it is possible to improve the K^* algorithm to maintain these critical provable guarantees. As a result, systems where both binding partners in the protein complex are flexible or mutable during the search can be accurately studied using K^*.

Intermutation pruning uses computed partition functions to truncate the conformation enumeration process for candidate sequences when they will provably fail to achieve a K^* score close to the best K^* score. To show that an intermutation pruning criterion [41] exists for PPI design we seek a halting condition for the conformation enumeration such that we know we have an ε-approximation to the

bound partition function for a given protein complex. First we observe: $K_i^* \geq \gamma K_0^*$, where K_i^* is the K^* score of the current sequence, K_0^* is the best score observed so far, and γ is a user-selected parameter. In the following lemma, n is the number of conformations in the search that remain to be computed, k is the number of conformations that have been pruned from the search with DEE, E_0 is the lower energy bound on all pruned conformations, R is the universal gas constant, and T is the temperature. The full partition function for the protein-protein complex, and unbound proteins are q_{AB}, q_A, and q_B respectively, while q_{AB}^*, q_A^*, and q_B^* denote the current calculated value of the partition functions during the computational search.

Lemma 1. *If the lower bound E_t on the minimized energy of the $(m+1)^{th}$ conformation returned by A^* satisfies $E_t \geq -RT(\ln(\gamma\varepsilon K_0^* q_A^* q_B^* - k\exp(-E_0/RT)) - \ln n)$, then the partition function computation can be halted, with q_{AB}^* guaranteed to be an ε-approximation to the true partition function, q_{AB}, for a candidate sequence whose score K_i^* satisfies $K_i^* \geq \gamma K_0^*$.*

This lemma shows that even when designing for protein-protein interactions, there exists a sequence pruning criterion during the K^* search.

Now we show that we can obtain a provable guarantee on the accuracy of the K^* score for each protein conformation. Since both partition functions are ε-approximations, we no longer obtain an ε-approximation to the K^* score but rather the following:

Lemma 2. *When amino acid substitutions (or flexible residues) are allowed on both strands in the computational design, the computed K^* score is a σ-approximation to the actual K^* score, where $\sigma = \varepsilon(2-\varepsilon)$.*

Since neither of the protein complex partition functions are calculated fully, the K^* score approximation is a 2ε-approximation as opposed to the ε-approximation for small molecule designs. This implies that we must compute better partition function approximations than before to maintain the same level of K^* score approximation. Nevertheless, the fact that the K^* score can still be provably approximated, confers all the advantages of a provable algorithm as stated above. The proofs of Lemmas 1 and 2 are provided in Text S1.

Computational Designs with K^*

The previously-determined NMR structure of the CAL PDZ domain bound to the C-terminus of CFTR (PDB ID: 2LOB) was used to model the binding of CAL to CFTR. To prepare the protein complex for the computational design, the initial complex structure was obtained by molecular dynamics refinement of the NMR structure as described previously [52]. Hydrogens were added to the structure using Reduce [53]. The CFTR peptide in the NMR structure was truncated to the six most C-terminal amino acids. An acetyl group was modeled onto the N-terminus of the peptide using restrained molecular dynamics and minimization in which the N-terminus of the peptide was allowed to move, while the remainder of the protein complex was restrained using a harmonic potential [54]. The coordinates of this starting structure are provided as supporting information (Text S2).

An 8 Å shell around the peptide hexamer was used as the input structure to K^*. The CFTR C-terminal residues, VQDTRL, were mutated to the following residues during the design search: P^{-5} to W, P^{-4} stayed fixed to Q, P^{-3} to all amino acids except Pro, P^{-2} to T/S, P^{-1} to all amino acids except Pro, and P^0 to I/L/V. In addition, the Probe program [55] was used to determine the side-chains on CAL that interact with the CFTR peptide mimic. The nine residues that interact with the peptide, as well as the two most N-terminal residues on the peptide, were allowed to be flexible during the design search (Fig. 1A). To explore the feasibility of our new algorithms, unless otherwise noted, full partition functions were not computed and a maximum of 10^3 conformations were allowed to contribute to each partition function.

Rotamer values were taken from the Penultimate Rotamer Library modal values [14]. The energy function used to evaluate protein conformations has been previously described [40,42]. The energy function, $E = \mathrm{vdW} + \mathrm{Coul} + \mathrm{EEF1}$, consists of a van der Waals term, a Coulombic electrostatics term, and an EEF1 implicit solvation term [56]. The EEF1 solvation term implicitly models water solvent during all of the computational designs. All design runs used the Amber98 [57] forcefield terms except for one prospective design run which used the Charmm19 [58] forcefield parameters.

Training of Energy Function Weights

Previously-determined experimental binding constants [59] for 16 of CAL's natural ligands were used to train the energy function weight parameters (See Text S1 Section S2). K^* scores were computed for each of the natural ligands. For this training, the CAL-CFTR structure only included the four most C-terminal residues of the peptide inhibitor. A gradient descent method was used to optimize the correlation between the K^* scores and the experimental K_i^{-1} values. The final parameters chosen for the design runs are as follows: a van der Waals scaling of 0.9, a dielectric constant of 20, and a solvation scaling of 0.76.

Peptide Array Comparison

K^* was used to predict binding between the CAL PDZ domain and the HumLib set of 6223 human protein C-termini. The binding of the C-termini peptides to CAL was experimentally assessed using a peptide SPOT array [59,60]. Due to experimental restrictions, all cysteines in the HumLib peptide set were replaced by serine in the peptide array. For consistency, all computational predictions compared to the array modeled serines in the place of cysteines. A summary of the peptide array data is presented in Fig. 3 while the complete binding results from the array are provided as **Supporting Information** (Table S1). The K^* algorithm was used to evaluate 4-mer structural models of 6223 peptide-array sequences to verify the accuracy of the algorithm's predictions. To compare the array data with the K^* predictions, the quantitative array data, measured in biochemical light units (BLUs), was converted into a binary yes/no CAL binding event. In other words, by using a fixed cutoff value, each sequence from the array was classified as either a CAL binder or non-binder. The cutoff value was chosen as three standard deviations away from the average BLU value of the array. A receiver operating curve (ROC), which uses a floating cutoff to compare array data to K^* scores, was used to evaluate the ability of K^* to predict the array binding data.

After the K^* predictions were calculated, the binding of C-termini peptides to CAL was also experimentally assessed using an additional SPOT array. The profile library array (ProLib; Fig. S3 in Text S1) was designed based on the following motif: bbbb $B_{-3}B_{-2}B_{-1}B_0$ (B = permutation of a defined set of amino acids, b = mixture of 17 amino acids, without C, M and W). The defined set of amino acids were selected based on the HumLib results combined with substitutional analyses [60] with B_{-3} = A/C/D/E/F/I/K/L/M/N/Q/R/S/T/V/W/Y, B_{-2} = S/T, B_{-1} = A/C/D/E/F/I/K/L/M/N/Q/R/S/T/V/W/Y, B_0 = I/L/V (Total number of peptides = 1734+22 internal control sequences). Incubation condition: 10μg/ml His-tagged CAL PDZ domain detected by anti-His (Sigma; 1:2600)/anti-mouse-HRP (Calbiochem; 1:2000) antibody sandwich.

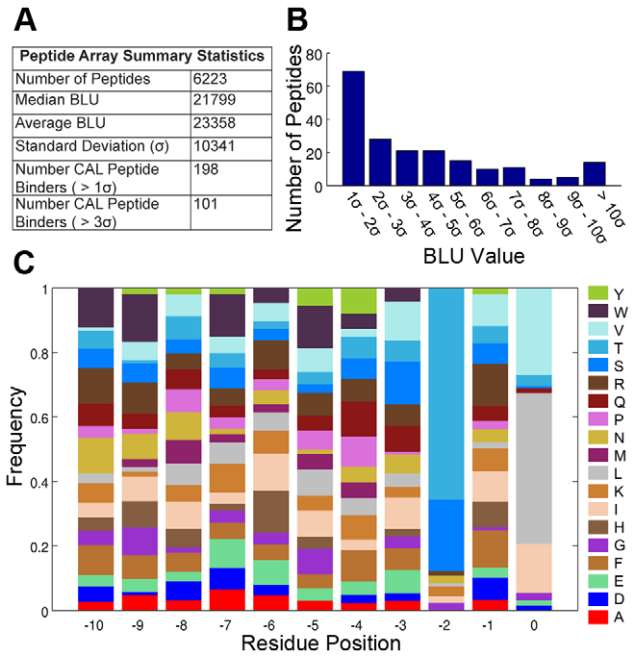

Figure 3. Summary of CAL peptide array. (A) Summary statistics for peptide array. Higher BLU (biochemical light unit) values indicate stronger protein binding to a peptide. (B) Distribution of the peptide BLU values from the peptide array in units of standard deviation above the mean (σ). (C) Normalized amino acid frequencies for the top sequences that have a BLU value greater than 3 standard deviations from the average, which were considered as the peptides that bound CAL for the validation of K^* predictions. The frequency of each amino acid type for each residue position was normalized by the total number of occurrences of that amino acid in the array at the given residue position.
doi:10.1371/journal.pcbi.1002477.g003

Prospective Computational Predictions

K^* was used to search over all peptide sequences within the CAL PDZ domain sequence motif (excluding prolines) to find new CAL peptide inhibitors. For computational efficiency the number of conformations enumerated by A* for each partition function was limited to 10^3 conformations. Two sets of peptides (promising designs and poorly ranked designs) were chosen to be experimentally validated.

In order to choose the most promising peptide inhibitors, a second K^* design was done where K^* scores for the top 30 sequences were re-calculated with the number of enumerated conformations per partition function increased to 10^5. Several top-ranked sequences were chosen to be experimentally tested. First, the top 7 ranked sequences from the second run were chosen. In addition, two sequences that greatly increased in ranking from the first to second run (rank 29 to 9, and rank 28 to 11) were chosen as well. Finally, a K^* run was conducted using Charmm forcefield parameters instead of Amber parameters. Two sequences that scored high on both the Amber and Charmm runs were chosen to be experimentally tested as well (Table 1).

The poorly-ranked designs were chosen to minimize the sequence similarity among the set of poorly-ranked peptides (Table 2). First, the worst-ranked peptide was chosen and added to initialize the set of negative sequences. Next, sequences were successively chosen from the worst 200 K^* ranked sequences and added to the set in order to maximize the amino acid sequence diversity with all the sequences already in the set. The similarity between two sequences was determined using the PAM-30 similarity matrix [61]. In total 23 (eleven top-ranked and twelve poorly-ranked) K^*-computed peptide inhibitor sequences were experimentally tested.

Table 1. Experimental validation of top-ranked K^* predictions.

Name	Sequence	K^* Ranking (out of 2166)	Experimental K_i (μM)
kCAL01	Ac-WQVTRV	9	$2.3 \pm 0.2^{\dagger}$
kCAL02	Ac-WQFTRL	$1^{\ddagger}$	$7.6 \pm 0.7^{\dagger}$
kCAL03	Ac-WQKTRL	2	$9.0 \pm 0.6^{\dagger}$
kCAL04	Ac-WQRTRL	5	$10.8 \pm 0.7^{\dagger}$
kCAL05	Ac-WQKTRI	4	$12.0 \pm 0.9^{\dagger}$
kCAL06	Ac-WQKTRV	1	16 ± 2
kCAL07	Ac-WQFTKL	$2^{\ddagger}$	16 ± 1
kCAL08	Ac-WQRTRI	7	16 ± 2
kCAL09	Ac-WQLTKL	11	17 ± 1
kCAL10	Ac-WQKTKL	6	17.8 ± 0.8
kCAL11	Ac-WQRTRV	3	18 ± 1

†K_i values with a binding affinity higher than the best previously known hexamer ($14 \pm 1 \mu$M). These sequences are shown in green in Fig. 5.

‡Sequence rank obtained by ordering the quantity: $\frac{R_A + R_C}{2}$, where R_A is the sequence rank from a design run using the Amber forcefield and R_C is the sequence rank from a run using the Charmm forcefield.
doi:10.1371/journal.pcbi.1002477.t001

Measuring Peptide Inhibitor Constants

The inhibitor dissociation constants of top- and poorly-ranked peptide sequences from the K^* CAL-CFTR design were experimentally determined. As a control, the best known peptide hexamer was also retested. The corresponding N-terminally acetylated peptides were purchased from NEO BioScience (Cambridge, MA) and the K_i values for the peptides were detected using fluorescence polarization (FP), using the method previously described in [59]. Briefly, the CAL PDZ domain was incubated in

Table 2. Experimental validation of poorly-ranked K^* predictions.

Name	Sequence	K^* Ranking (out of 2166)	Experimental K_i (μM)
kCAL20	Ac-WQYTMI	1981	24 ± 4
kCAL21	Ac-WQYTDL	2082	32 ± 4
kCAL22	Ac-WQISWL	1973	37 ± 15
kCAL24	Ac-WQHTEV	1989	87 ± 7
kCAL23	Ac-WQMTDI	1969	90 ± 9
kCAL25	Ac-WQCSEI	2051	107 ± 9
kCAL26	Ac-WQESEL	2095	120 ± 20
kCAL27	Ac-WQDTWI	2158	400 ± 20
kCAL28	Ac-WQWSDV	2166	400 ± 200
kCAL29	Ac-WQDSCV	2011	1000 ± 200
kCAL30	Ac-WQGSDV	2075	2200 ± 300
kCAL31	Ac-WQDSGI	1992	> 5000

doi:10.1371/journal.pcbi.1002477.t002

FP buffer (25 mM Tris-HCl pH 8.5, 150 mM NaCl; supplemented to a final concentration of 0.1 mg/mL bovine IgG (Sigma) and 0.5 mM Thesit (Fluka)) with a labeled peptide of known binding affinity. Each peptide inhibitor was serially diluted and the protein-peptide mixture was added to each dilution. Finally, the amount of competitive inhibition was tracked using residual fluorescence polarization at temperatures between $25-28^{\circ}C$. Each K_i value is reported as an average of three FP experiments conducted on separate days along with the corresponding standard deviation.

Measuring Chloride Flux

Ussing chamber experiments were performed as described previously [11]. Polarized monolayers of patient-derived bronchial epithelial cells, CFBE-ΔF cells (a generous gift of Dr. J.P. Clancy [62,63]), were maintained in MEM with 2 mM l-glutamine, 10% fetal bovine serum, 50 units/mL penicillin, $50\mu g/mL$ streptomycin, $2\mu g/mL$ puromycin, $5\mu g/mL$ plasmocin, and $2.5\mu g/mL$ amphotericin B. Cells were grown at $37^{\circ}C$ in 5% CO_2. Twenty four hours before treatment the cells were moved to MEM with only penicillin and streptomycin. Peptides were dissolved in DMSO and diluted to $500\mu M$ in PBS. Peptide solutions were applied to cells following incubation with BioPORTER delivery reagent (Sigma). The final DMSO concentration did not exceed 0.03%. Following a 3.5 hour incubation with peptide, short circuit currents (I_{SC}) were monitored in Ussing chambers. Following treatment with amiloride, forskolin, and genistein, ΔF508-CFTR chloride flux was measured as the change in I_{SC} when the CFTR-specific inhibitor, $CFTR_{inh}172$ [64,65], was applied to the cell monolayer. All measurements were performed at $37^{\circ}C$.

Results

We applied the K^* algorithm to the CAL-CFTR system to find a CAL PDZ peptide inhibitor that acts as a biologically active stabilizer of ΔF508-CFTR. First, we developed the ensemble-based computational structural design software K^* to design PPIs. To validate the design methodology, the predictions of the K^* algorithm were compared with binding data of CAL binding human protein C-termini. The validation showed K^* was able to enrich for peptide inhibitors. We then used K^* to prospectively find new peptide inhibitors of CAL. The top-scoring predicted sequences were experimentally validated and we determined that they all bind CAL with μM affinity. Next, additional binding data for peptide sequences that match the known CAL binding motif were collected and compared to the K^* predictions. Finally, Ussing chamber experiments showed that the highest affinity designed peptide significantly rescues ΔF508-CFTR in bronchial epithelial cells.

Validation of the K^* Algorithm

To validate the K^* algorithm, we compared K^* predictions for CAL peptide inhibitors against peptide array binding data. First, peptides from the 6223 peptide HumLib library were tested for CAL binding using a SPOT array [59]. The array was able to find over one hundred peptides that clearly bind the CAL PDZ domain (Fig. 3). Second, K^* predictions were made for all of the peptide sequences in the HumLib library. Fig. 4A shows the resulting receiver operating curve (ROC) when comparing the K^* scores to the binding measurements (BLU values) of the peptide array. The ROC has an area under the curve (AUC) of 0.84 which shows that K^* greatly enriches for peptides that bind CAL. Specifically, according to the peptide array, out of the top 30 K^* predicted sequences, 11 are expected to bind CAL. Notably, this is a 20-fold

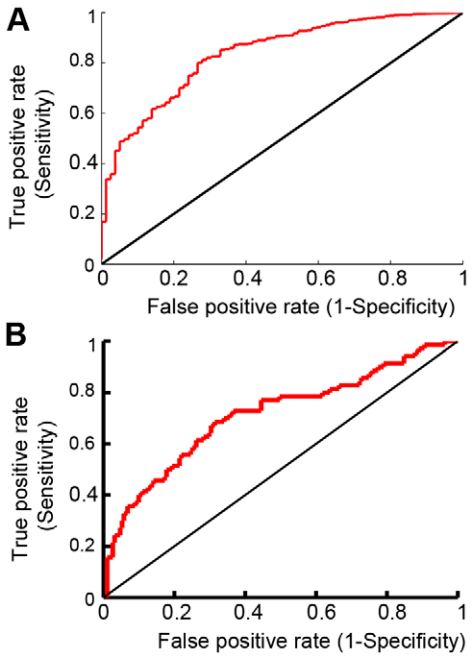

Figure 4. K^* enriched for peptide sequences that bind the CAL PDZ domain. ROCs were calculated comparing K^* predictions to (A) the entire HumLib peptide array data set (AUC = 0.84) and (B) only sequences in the HumLib array that matched the CAL binding motif (AUC = 0.71).
doi:10.1371/journal.pcbi.1002477.g004

increase over the number of binders that would be expected to be found if the CAL binding peptides were distributed randomly within the K^* predictions.

To investigate the success of the algorithm in more detail, we evaluated the importance of the CAL binding motif in determining K^* predictions. The amino acid frequencies from the top binding peptides of the HumLib library (Fig. 3C) and natural binding partners of CAL [59] reveal that the canonical sequence motif of CAL is X-S/T-X-L/V/I. As expected, among the full set of HumLib peptides, K^* enriches for sequences that conform to this motif. Furthermore, if we allow K^* to design peptides varying at the primary motif positions 0 and −2, it achieves an AUC of 0.94 (Text S1 Section S3 and Fig. S2 in Text S1), confirming its ability to identify the motif *de novo*. While K^* also identified a few non-motif sequences in each case, the HumLib suggests that CAL actually can bind to such sequences, albeit less frequently (10 of 5867 sequences).

Of course, the identification of motif residues, while a necessary test of the algorithm, does not by itself represent a major advance in affinity prediction. The HumLib library shows that only 70 out of 261 sequences with the CAL binding motif bind to CAL. A much more stringent test of the K^* design algorithm is thus to determine how well K^* enriches for binders among sequences that match the known CAL binding motif. As a first test, we recalculated the ROC curve considering only peptides in the HumLib library that match the CAL sequence motif, and K^* was still able to significantly enrich for CAL peptide binders (AUC = 0.71; Fig. 4B). This search, together with the blind test of K^* rankings described below, provides a true test that the success of K^* in predicting HumLib binders is not merely due to its identification of peptides conforming to the known sequence motif, but also to its ability to distinguish high- and low-affinity binders among such peptides.

Prospective Design of CAL Peptide Inhibitors

While SPOT arrays have proven to be a powerful tool for the identification of CAL binding peptides, the highest affinity inhibitors identified to date are composed of at least 10 amino acids. For hexamers, the highest published affinity is for iCAL35 (WQTSII; [60]). Since K^* was able to successfully enrich for CAL binders found in the HumLib library, we then used K^* to prospectively find novel, shorter CAL peptide inhibitors, searching over 2166 peptides containing motif-based combinations of the C-terminal four residues. To facilitate accurate experimental binding-constant measurements, each peptide was extended by a shared N-terminal addition of the most frequent P^{-5} and P^{-4} residues among HumLib binders(WQ), yielding hexamer sequences that exhibit a higher baseline affinity [59]. Both top- and bottom-ranked sequences were chosen for experimental validation. The K_i value for each peptide hexamer was determined using fluorescence polarization [59] (Table 1). We used the same FP protocol to confirm the affinity of the acetylated iCAL35 reference peptide for CAL ($K_i = 14 \pm 1 \mu M$).

All of our top-ranked inhibitors are novel CAL ligands, for which neither predicted nor experimental affinities were previously available. Remarkably, all of the top predicted peptides bind CAL with high affinity (Fig. 5A, Table 1). The tightest binding predicted peptide (kCAL01, WQVTRV) had a K_i of $2.3 \pm 0.2 \mu M$. While this affinity is comparable to that of several other PDZ inhibitors [66,67], solution-state measurements show that the CAL PDZ domain exhibits systematically weak interactions with target C-termini: note that the K_i for the wild-type CFTR sequence (TEEEVQDTRL) is $390 \mu M$ and the best known affinity natural ligand (ANGLMQTSKL) for CAL is $21 \mu M$ [60]. Thus, our design algorithm successfully identifies high affinity peptide inhibitors of the CAL PDZ domain, with 170-fold higher affinity than the interaction we were trying to inhibit and 9-fold higher affinity than any comparable natural ligand. This peptide affinity advantage may be important in physiological applications, since the native CAL:CFTR target interaction may involve additional sources of affinity outside the PDZ binding pocket [4,59], not available to a peptide inhibitor.

We also performed further analysis of the HumLib SPOT array used for K^* validation. Selecting the most common amino acid at positions P^0 to P^{-5} among HumLib binders yields the sequence WQSTRL (HumLib01, Fig. 3C), which is ranked in the top 50 K^* predictions (out of 2166). This sequence is also the strongest binder identified among the ProLib sequences (see below, and Fig. S3 in Text S1). However, when we measured the CAL binding for HumLib01 using fluorescence polarization (FP) it exhibited a K_i value of $13.5 \pm 0.5 \mu M$, only a marginal improvement in affinity compared to iCAL35 ($14 \pm 1 \mu M$). In comparison, five of the eleven top K^* predicted sequences we measured with FP show an improvement in binding compared to both iCAL35 and HumLib01, and kCAL01 shows a six-fold improvement over both iCAL35 and the HumLib01 sequence.

The best inhibitor found through previous FP and array screens involves a fluorescein group modification to a peptide decamer (F^*-iCAL36, F^*-ANSRWPTSII, $K_d = 1.3 \mu M$). kCAL01 rivals this binding affinity despite the computational search library restriction to only allow amino acids and hexamer sequences. Critically, at 830 Da, kCAL01 has approximately twice the binding efficiency (ratio of inhibitor potency, ΔG, to molecular mass) of F^*-iCAL36 and is much closer in size to typical drugs. This makes kCAL01 a very promising inhibitor compared to F^*-iCAL36 and other discovered inhibitors.

Furthermore, as suggested by our retrospective tests, the tight binding of our top-ranked sequences was not merely a consequence of the underlying CAL-binding motif used to select candidate sequences for evaluation. To establish this, we selected a set of poorly-ranked peptides to minimize sequence similarity and evaluated their CAL-binding affinity experimentally. Almost all of the poorly-ranked sequences bound CAL, consistent with their motifs (Fig. 5A). Reflecting the enrichment of CAL binders in the pool, the two poorly-ranked peptides with the best affinities ($K_i = 24 \mu M$ and $32 \mu M$, respectively) were indeed close to the affinity of the weakest top-ranked sequence ($K_i = 18 \mu M$). However, all of the poorly ranked peptides bound CAL more weakly than any of the top-ranked sequences (Table 1), and none of them had

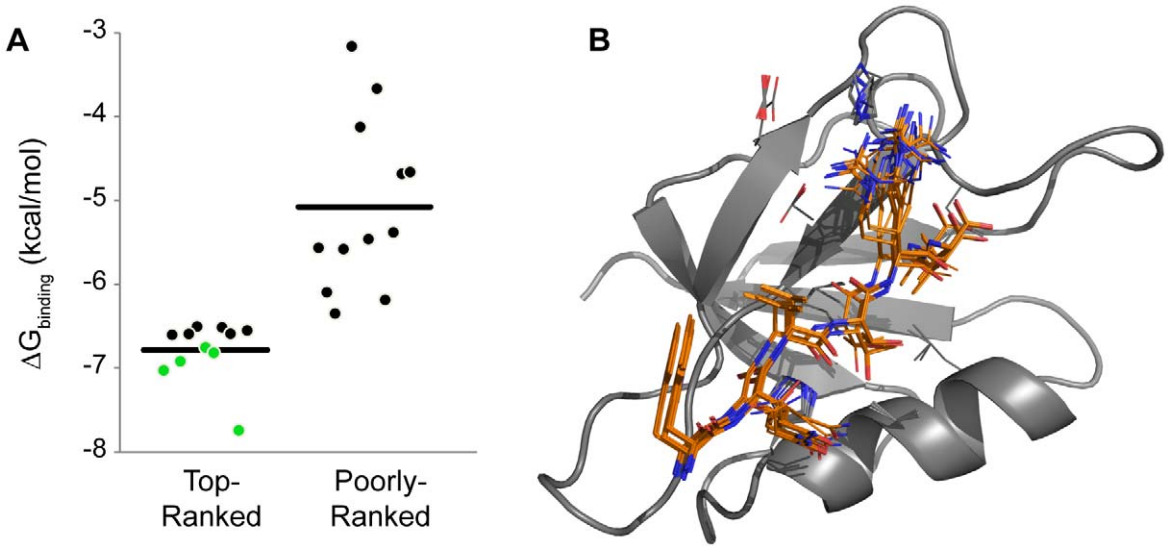

Figure 5. (A) ΔG values for top- and poorly-ranked K^* predictions that were experimentally tested using fluorescence polarization. Predictions plotted in green denote that the binding affinity was higher than the best previously known hexamer (14 μM). Horizontal line represents average ΔG for plotted sequences. Sequence information and binding data can be found in Tables 1 and 2. (B) Ensemble of top 100 conformations for the peptide (kCAL01: WQVTRV, orange sticks) with tightest binding to CAL (gray ribbon).
doi:10.1371/journal.pcbi.1002477.g005

improved affinity relative to prior biochemical efforts. This suggests that K^* can efficiently distinguish among motif-bearing peptides, allowing it to predict sequences with CAL affinities unprecedented among hexamers.

Detailed analysis of the K^* predictions suggests that the use of both ensemble-weighting and minDEE approaches was important in the success of the algorithm. The ensembles generated by K^* do not have a dominant conformation, i.e., a conformation with significantly lower energy than the others, which would thus dominate in the partition function. For example, in the case of iCAL35 (WQTSII), K^* found 75 conformations that were within 0.5 kcal/mol and 454 conformations that were within 1 kcal/mol of the iCAL35 GMEC. In general, the ensemble conformations are consistent with canonical PDZ:peptide interactions and with the conformation of the CAL-bound CFTR peptide determined by NMR [52]. To determine the importance of the ensemble-based K^* rankings we compared the predictions to two single-structure GMEC-based methods, minDEE [41], and rigid-rotamer DEE (rigidDEE) [68]. Both minDEE and rigidDEE were run with the same energy parameters as the K^* designs. However, since the single-structure designs only compute the energy of the bound state, reference energies [16] were included as in [69] to account for the energy of the unbound state. The inclusion of reference energies for single-structure designs have been deemed necessary by most protein designers to account for the unfolded/unbound state [24,69,70]. K^* does not need reference energies since it calculates a partition function for both the bound and unbound states of the complex [16,40]. Therefore, reference energies are included to make the comparison between K^* and the single-structure designs more fair. We compared the top 30 sequences from minDEE and rigidDEE and found they had no sequences in common. This supports previous work where we have shown that in over 69 protein design systems minDEE finds low energy sequences that rigidDEE discards by not allowing minimization [41,50]. In addition, when we compare the top 30 rigidDEE and minDEE results to the top K^* designs we find that they have only three and four sequences in common, respectively. If we had used only GMEC-based approaches instead of K^*, we would not have predicted most of the experimentally successful sequences that K^* found, including the best inhibitor kCAL01. In addition, the overall sequence rankings show a very poor correlation between the minDEE and K^* predictions; the same is true of the rigidDEE and K^* predictions ($R^2 = 0.1$ and 0.09 respectively).

Blind Test of K^* Predictions within the CAL Binding Motif

The prospective peptide predictions demonstrate that K^* can successfully find CAL peptide inhibitors. Our solution-state binding tests provide robust information for the best and worst K^*-predicted peptides, but give little information about the CAL binding of the remaining peptides that match the CAL motif. To investigate this experimentally, we designed a peptide library SPOT array (ProLib) based on the HumLib motif combined with substitutional analyses [60]. The resulting sequences closely match our prospective prediction set and the binding of these sequences to CAL was assessed as described in the Materials and Methods section. Using a similar analysis to that performed on the HumLib peptide array we compared the K^* predictions to the CAL binding observed with the ProLib array. We found an AUC = 0.88 (Fig. 6). Note that this AUC is much higher than the 0.71 found when only looking at CAL motif sequences within the HumLib array. One explanation for this improvement is that the experimental setup is closer to the design model used by K^*. Specifically, the ProLib array uses a mixture of amino acids at P^{-4} to P^{-7} of the peptides,

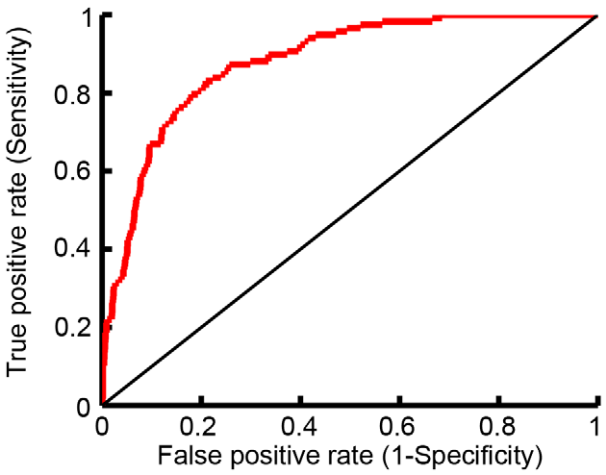

Figure 6. K^* was used to predict binding between the CAL PDZ domain and the peptide array, ProLib (Figure S3), which contained peptide sequences that match the CAL binding motif. The ROC curve shown compares the K^* predictions to the observed peptide array binding data. AUC = 0.88.
doi:10.1371/journal.pcbi.1002477.g006

while the HumLib array is composed of decamer peptides. Thus, the ProLib data focuses on the identity of the last 4 C-terminal positions, which better matches the sequence and structure search space of the K^* designs. A complete evaluation of the accuracy of K^* affinity predictions would require the synthesis and FP binding analysis of all 2166 sequences within the CAL binding motif. However, taken together, the FP measurements for the designed peptides plus the ProLib blind test suggest that K^* is a powerful filter, efficiently selecting tight binders from a pool of sequences with baseline affinity for the target.

Biological Activity of the Highest Affinity Designed Peptide Inhibitor

All of our top-predicted inhibitors successfully bound CAL, which suggests that they should disrupt the degradation pathway of CFTR. The ability of kCAL01 to restore ΔF508-CFTR function was assessed by measuring CFTR-mediated chloride efflux in CF-patient derived bronchial cells expressing ΔF508-CFTR (CFBE-ΔF) using an Ussing chamber apparatus [11]. As a control peptide, we used kCAL31 (WQDSGI), which was ranked as the weakest interactor by K^* and for which no binding was detected experimentally (Table 2). Fig. 7 shows ΔF508-CFTR chloride secretion across polarized monolayers treated with either kCAL31, the iCAL35 reference peptide, or kCAL01. Previous studies with fluorescently labeled peptides have demonstrated delivery into CFBE-ΔF cells using the BioPORTER reagent [11]. Significance of rescue was evaluated by comparing percentage improvement in chloride efflux to rescue from a well-established "corrector" under identical conditions, and by Student's t-test (p-value). Compared to the non-binding control, the previously best hexamer, iCAL35, yields only a slight (non-significant) improvement in chloride secretion (4%, $p = 0.16$). In contrast, chloride secretion following treatment with the designed inhibitor kCAL01 is significantly enhanced with respect to the control peptide (12%, $p = 0.0049$) and with respect to the reference (8%, $p = 0.037$) peptide. Indeed, the biological activity of kCAL01 is very similar to that observed under similar conditions following treatment with either the best previously available CAL inhibitor (F^*-iCAL36) or the first-generation corrector corr-4a [6,11].

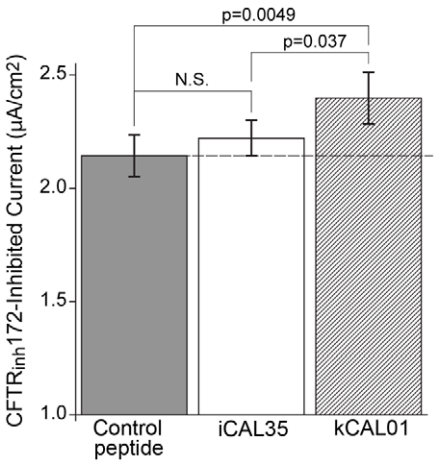

Figure 7. Top binding peptide is biologically active. The ΔF508-CFTR specific chloride flux is shown for a control peptide (kCAL31; WQDSGI; no CAL binding detected), the reference peptide (iCAL35; WQTSII), and the tightest binding design peptide (kCAL01; WQVTRV). kCAL01 shows a 12% increase in chloride efflux over the control peptide. p values shown are for pairwise comparisons ($n = 12$). Values shown are mean $\pm$ standard error of the mean (SEM). N.S.: *not significant*, $p = 0.16$.
doi:10.1371/journal.pcbi.1002477.g007

Discussion

The new K^* algorithm has enabled the design of the first high-affinity hexapeptide CAL PDZ inhibitor with demonstrated ability to rescue ΔF508-CFTR. By interfering with CAL-mediated degradation, our best designed peptide, kCAL01, can act as a CFTR "stabilizer," allowing ΔF508-CFTR to recycle back into the membrane. Currently the only well-studied ways to rescue mutant CFTR function with drug-like molecules are through "potentiators" and "correctors" which do not address the problem that ΔF508-CFTR is rapidly endocytosed and degraded at physiological temperatures [9]. Like other CAL inhibitors, kCAL01 should work in conjunction with potentiators and correctors to create an additive effect [11].

kCAL01 was observed to increase ΔF508-CFTR activity by 12%. While this effect is clearly statistically significant ($p = 0.0049$), we also wished to assess its magnitude relative to the effect of known rescue compounds. The performance of kCAL01 was benchmarked using polarized human airway epithelial cells derived from a CF patient (stably expressing ΔF508-CFTR; CFBE-ΔF cells). In these cells, CFTR rescue is more challenging than in heterologous cells, but the levels of rescue observed are more likely to reflect the physiological situation. Since CFTR modulation is extremely sensitive to experimental conditions, and particularly to the type of cells used [8,71], we chose to compare the performance of kCAL01 against the corrector corr-4a. There are two reasons for this choice for comparison: (a) corr-4a is a well established benchmark for CFTR correctors [72]; and (b) directly comparable data are available based on our previous studies [1]. Under identical experimental conditions, corr-4a produces a 15% increase in ΔF508-CFTR levels in CFBE-ΔF cells [1]. Thus, the 12% increase seen with the kCAL01 inhibitor peptide is similar to that produced by a first-generation corrector. Since corr-4a and kCAL01 have orthogonal mechanisms of action, this enables additive rescue as an attractive treatment option. Specifically, in the long term the therapeutic impact of CAL inhibitors is likely to be enhanced by their ability to provide additive rescue with correctors, offering the prospect of combination treatment [11].

To design kCAL01 we developed a novel, provable, ensemble-based protein design algorithm for protein-peptide and protein-protein interactions. The validation of K^* by comparing its predicted binding scores to CAL peptide-array data demonstrates K^*'s strong ability to enrich for human protein sequences that bind CAL. While the HumLib array showed that CAL binds a specific motif, it also shows (along with the ProLib array) that CAL does not bind all sequences that match the motif. In HumLib, 191 of 261 sequences that match the motif did not bind CAL. Moreover, all of the peptides synthesized for this work (kCAL01-kCAL31) match the CAL motif, but have a wide range of binding affinities. Therefore, K^* needs to perform the difficult task of differentiating the affinities of peptides that share the CAL motif, rather than merely separating motif from non-motif sequences. The HumLib analysis, FP analysis of top and poorly-ranked K^* predictions, and the ProLib analysis all show that K^* is able to enrich for sequences within the CAL PDZ sequence motif that have high-affinity interactions with CAL.

The experimental validation of top-ranked K^* sequences confirms that K^* prospectively predicted novel high-affinity CAL peptide inhibitors. Compared to the inhibitory constant of the natural CFTR C-terminus, the designed sequences are much stronger binders. Indeed, our approach found peptide sequences that bound more tightly than iCAL35, the best previously known hexamer sequence. Interestingly, even though iCAL35 binds to the CAL PDZ domain, it is unable to mediate significant or substantial rescue of ΔF508-CFTR in CFBE-ΔF cells (Fig. 7). The designed inhibitor's improvement in binding directly translates to increased ΔF508-CFTR activity in CF-patient derived airway epithelial cells, demonstrating the value of using our computational approach to design protein:peptide interactions.

Current therapeutics known to rescue CFTR function are small molecules generally discovered through high throughput library screens [72]. To find CFTR stabilizers we needed to discover inhibitors that could block the CAL-CFTR PPI. Unfortunately, small molecules that inhibit PPIs are rare and the development of such inhibitors has been very difficult due to the shallow, distributed nature of the interfaces [73]. Therefore, we have focused on tools to design peptide inhibitors, developing and validating a new K^* algorithm that has identified low molecular weight, high-affinity sequences. While our previous work employed high-throughput peptide arrays to screen for inhibitors [60], the computational design approach can easily and accurately be expanded beyond the limits of peptide array synthesis, providing a novel avenue for identifying CF therapeutic leads with improved affinity, specificity, and proteolytic stability.

In this paper we have focused on improving peptide inhibitor affinities, but our success suggests that K^* can also be used to improve peptide specificity and proteolytic stability. For optimal biological efficacy, CAL inhibitors should avoid off-target effects, including interactions with other CFTR trafficking proteins (Fig 1B), such as the NHERF family [3]. To achieve peptide specificity, K^* could be run to find peptides that did not bind well to these off-target interactors, a process known as *negative design* [16,42]. The experimentally-tested poorly-ranked K^* predictions all had a worse affinity for CAL than the top-predicted peptides (Tables 1 and 2). This suggests that K^* has the capability to conduct negative design for the CAL system. Also, we have shown the successful application of K^* negative design to other biological systems [42]. Finally, since the efficacy of natural peptides is often limited by proteolytic stability, it could be beneficial to extend the K^* software to incorporate non-natural amino acids, such as d-amino acids, into the design search space. This will allow the design of compounds that inhibit CAL, but cannot be degraded as readily as linear L-peptides.

The K^* scoring function uses energy terms for electrostatics, van der Waals energy, and implicit solvation. K^* also utilizes an approximation of conformational entropy factors through its ensemble-based scoring [16,41]. Analysis of these components can potentially identify important interactions in the top peptide inhibitor designs. Comparing the average energy contribution for the top 30 predictions to the median for all designs we find that all components contribute favorably to the peptide binding, with van der Waals giving the largest benefit (−11.2 kcal/mol), followed by electrostatics (−10.9 kcal/mol), and finally solvation (−8.2 kcal/mol). However, even within the top 30 predictions the dominant energetic component varies greatly (electrostatics is dominant for 12 sequences, van der Waals for 6 sequences, and solvation for 12 sequences).

Tidor and co-workers [69] have suggested that design predictions are best when re-ranking structures using a purely electrostatic energy function. We addressed this possibility by comparing the AUC obtained from a purely electrostatic function vs. that obtained from our complete energy function. If we use only the electrostatic term, the AUC was 0.61 (bound energy only) or 0.66 (bound minus unbound). Both values are significantly lower than the 0.84 AUC value obtained with the full function. Thus, while electrostatic terms are important to the success of the algorithm, inclusion of a more complete energetic model improves the prediction. In fact, no individual energy term outperforms the K^* score when classifying the peptide array data. Thus, K^* predicts its successful designs by accurately incorporating all three energy terms through ensemble-based scoring.

Many of the binding sequences identified by K^* contain a positively charged residue (R/K) at P^{-1}. Similarly, in the HumLib array, about 26% of the sequences that we consider to be binders contain a positively charged residue at P^{-1}, and in the ProLib array 53% of the binders contain an R/K at P^{-1}. Based on our previous NMR analysis [52], the P^{-1} Arg can form a salt-bridge with Glu309 on the periphery of the CAL binding site (Fig. 1A), an electrostatic contribution that could theoretically dominate the ROC curve analysis. However, because 74% of the top binding sequences in the HumLib array do not contain the P^{-1} R/K, the strong K^* AUC values suggest that it must also correctly predict these sequences. To test this assertion more forcefully, we removed all of the sequences with a positively charged residue at position −1 and then recalculated the ROC curve. This results in an AUC of 0.82, almost identical to the value of 0.84 obtained with all sequences. Thus, consistent with the significant contributions of each term in the energy function, the ROC behavior of the algorithm is not dependent on the presence or absence of a positively charged residue at P^{-1}.

A small number of K_i values were used to train the new K^* algorithm to properly scale energy terms for protein-peptide interactions, which can now be used for additional protein-peptide interaction designs. Besides the training, the only system specific data used was the input starting structure and CAL sequence motif. The sequence motif was used as an optional filter to expedite the search, but should not affect the ability of K^* to find high-affinity inhibitors. As seen from the HumLib peptide array comparison, K^* yields a higher ROC AUC when considering the entire array, which implies that K^* is better at distinguishing CAL peptide inhibitors from the entire sequence space than from within only the known sequence motif. This suggests K^* will be able to find new high-affinity inhibitors if the search space is expanded.

Beyond its utility in the design of enhanced CAL inhibitors, the K^* algorithm represents a general framework for analyzing PDZ domains and other protein-protein interfaces. PDZ domains are among the most common interaction domains in the human genome [74]. Using traditional biochemical approaches, the characterization of the binding affinity of candidate partners, as well as the identification of high-affinity reporters and inhibitors, often requires the individual synthesis of dozens of peptides, many of which fail to interact robustly. As shown for CAL, K^* offers a facile mechanism to predict affinities and to design novel ligand sequences using only an initial input structure. Furthermore, the proofs and algorithm presented here provide a general approach for modeling peptide-mediated PPIs that regulate a wide variety of critical physiological processes.

Availability

The source code of our program is freely available, and is distributed open-source under the GNU Lesser General Public License (Gnu, 2002). The source code can be freely downloaded at http://www.cs.duke.edu/donaldlab/osprey.php.

Supporting Information

Table S1 Binding data from CAL HumLib peptide array.
(PDF)

Text S1 Proof of Lemma 1 and 2. Additional methods detailing training of energy function weights and computational design of CAL motif residue positions.
(PDF)

Text S2 Structural coordinates for the K^* design starting template of the CAL PDZ domain:CFTR C-terminus complex.
(TXT)

Acknowledgments

The authors thank all members of the Donald Lab, in particular Mr. Pablo Gainza for helpful discussions and comments. We thank Mr. Lars Vouilleme for his critical reading of the manuscript.

Author Contributions

Conceived and designed the experiments: KER PB DRM BRD. Performed the experiments: KER PRC PB. Analyzed the data: KER PRC PB DRM BRD. Contributed reagents/materials/analysis tools: KER PRC PB DRM BRD. Wrote the paper: KER PRC PB DRM BRD.

References

1. Kim E, Sheng M (2004) PDZ domain proteins of synapses. Nat Rev Neurosci 5: 771–781.
2. Humbert P, Russell S, Richardson H (2003) Dlg, scribble and lgl in cell polarity, cell proliferation and cancer. Bioessays 25: 542–553.
3. Guggino WB, Stanton BA (2006) New insights into cystic fibrosis: molecular switches that regulate CFTR. Nat Rev Mol Cell Biol 7: 426–436.
4. Cheng J, Moyer BD, Milewski M, Loffing J, Ikeda M, et al. (2002) A golgi-associated PDZ domain protein modulates cystic fibrosis transmembrane regulator plasma membrane expression. J Biol Chem 277: 3520–3529.
5. Cheng J, Wang H, Guggino WB (2004) Modulation of mature cystic fibrosis transmembrane regulator protein by the PDZ domain protein CAL. J Biol Chem 279: 1892–1898.
6. Pedemonte N, Lukacs GL, Du K, Caci E, Zegarra-Moran O, et al. (2005) Small-molecule correctors of defective DeltaF508-CFTR cellular processing identified by high-throughput screening. J Clin Invest 115: 2564–2571.
7. Goor FV, Straley KS, Cao D, Gonzlez J, Hadida S, et al. (2006) Rescue of DeltaF508-CFTR trafficking and gating in human cystic fibrosis airway primary cultures by small molecules. Am J Physiol Lung Cell Mol Physiol 290: L1117–1130.
8. Rowe SM, Pyle LC, Jurkevante A, Varga K, Collawn J, et al. (2010) DeltaF508 CFTR processing correction and activity in polarized airway and non-airway cell monolayers. Pulm Pharmacol Ther 23: 268–278.
9. Cholon DM, O'Neal WK, Randell SH, Riordan JR, Gentzsch M (2010) Modulation of endocytic trafficking and apical stability of CFTR in primary

human airway epithelial cultures. Am J Physiol Lung Cell Mol Physiol 298: L304–314.
10. Wolde M, Fellows A, Cheng J, Kivenson A, Coutermarsh B, et al. (2007) Targeting CAL as a negative regulator of F508-CFTR Cell-Surface expression. J Biol Chem 282: 8099–8109.
11. Cushing PR, Vouilleme L, Pellegrini M, Boisguerin P, Madden DR (2010) A stabilizing inuence: CAL PDZ inhibition extends the half-life of dF508-CFTR. Angew Chem Int Ed Engl 49: 9907–9911.
12. Dunbrack RL, Karplus M (1993) Backbone-dependent rotamer library for proteins application to side-chain prediction. J Mol Biol 230: 543–574.
13. Janin J, Wodak S (1978) Conformation of amino acid side-chains in proteins. J Mol Biol 125: 357–386.
14. Lovell SC, Word JM, Richardson JS, Richardson DC (2000) The penultimate rotamer library. Proteins 40: 389–408.
15. Ponder JW, Richards FM (1987) Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequences for different structural classes. J Mol Biol 193: 775–791.
16. Donald BR (2011) Algorithms in Structural Molecular Biology. Cambridge, MA: The MIT Press.
17. Dahiyat BI, Mayo SL (1996) Protein design automation. Protein Sci 5: 895–903.
18. Dahiyat BI, Mayo SL (1997) De novo protein design: Fully automated sequence selection. Science 278: 82–87.
19. Desjarlais JR, Handel TM (1995) De novo design of the hydrophobic cores of proteins. Protein Sci 4: 2006–2018.
20. Koehl P, Levitt M (1999) De novo protein design. I. In search of stability and specificity. J Mol Biol 293: 1161–1181.
21. Koehl P, Delarue M (1994) Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. J Mol Biol 239: 249–275.
22. Jones DT (1994) De novo protein design using pairwise potentials and a genetic algorithm. Protein Sci 3: 567–574.
23. Jiang X, Pistor E, Farid RS, Farid H (2000) A new approach to the design of uniquely folded thermally stable proteins. Protein Sci 9: 403–416.
24. Kuhlman B, Baker D (2000) Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci U S A 97: 10383–10388.
25. Lee C, Subbiah S (1991) Prediction of protein side-chain conformation by packing optimization. J Mol Biol 217: 373–388.
26. Fromer M, Yanover C (2008) A computational framework to empower probabilistic protein design. Bioinformatics 24: i214–222.
27. Yanover C, Weiss Y (2003) Approximate inference and protein-folding. In: S Becker ST, Obermayer K, eds. Advances in Neural Information Processing Systems 15. Cambridge, MA: MIT Press. pp 1457–1464.
28. Gordon DB, Mayo SL (1999) Branch-and-Terminate: a combinatorial optimization algorithm for protein design. Structure 7: 1089–1098.
29. Hong E, Lippow SM, Tidor B, Lozano-Pérez T (2009) Rotamer optimization for protein design through MAP estimation and problem-size reduction. J Comput Chem 30: 1923–1945.
30. Leach AR, Lemon AP (1998) Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. Proteins 33: 227–239.
31. Althaus E, Kohlbacher O, Lenhof H, Muller P (2002) A combinatorial approach to protein docking with exible side chains. J Comput Biol 9: 597–612.
32. Kingsford CL, Chazelle B, Singh M (2005) Solving and analyzing side-chain positioning problems using linear and integer programming. Bioinformatics 21: 1028–1039.
33. Leaver-Fay A, Kuhlman B, Snoeyink J (2005) An adaptive dynamic programming algorithm for the side chain placement problem. Pac Symp Biocomput 10: 16–27.
34. Desmet J, Maeyer MD, Hazes B, Lasters I (1992) The dead-end elimination theorem and its use in protein side-chain positioning. Nature 356: 539–542.
35. Gilson M, Given J, Bush B, McCammon J (1997) The statistical-thermodynamic basis for computation of binding affinities: a critical review. Biophys J 72: 1047–1069.
36. Allen BD, Nisthal A, Mayo SL (2010) Experimental library screening demonstrates the successful application of computational protein design to large structural ensembles. Proc Natl Acad Sci U S A 107: 19838–19843.
37. Berezovsky IN, Chen WW, Choi PJ, Shakhnovich EI (2005) Entropic stabilization of proteins and its proteomic consequences. PLoS Comput Biol 1: e47.
38. Kamisetty H, Ramanathan A, Bailey-Kellogg C, Langmead CJ (2011) Accounting for conformational entropy in predicting binding free energies of protein-protein interactions. Proteins 79: 444–462.
39. Zhang J, Liu JS (2006) On Side-Chain conformational entropy of proteins. PLoS Comput Biol 2: e168.
40. Chen C, Georgiev I, Anderson AC, Donald BR (2009) Computational structure-based redesign of enzyme activity. Proc Natl Acad Sci U S A 106: 3764–3769.
41. Georgiev I, Lilien RH, Donald BR (2008) The minimized dead-end elimination criterion and its application to protein redesign in a hybrid scoring and search algorithm for computing partition functions over molecular ensembles. J Comput Chem 29: 1527–1542.
42. Frey KM, Georgiev I, Donald BR, Anderson AC (2010) Predicting resistance mutations using protein design algorithms. Proc Natl Acad Sci U S A 107: 13707–13712.
43. Brannetti B, Helmer-Citterich M (2003) iSPOT: a web tool to infer the interaction specificity of families of protein modules. Nucleic Acids Res 31: 3709–3711.
44. Thomas J, Ramakrishnan N, Bailey-Kellogg C (2009) Graphical models of protein-protein interaction specificity from correlated mutations and interaction data. Proteins 76: 911–929.
45. Smith CA, Kortemme T (2010) Structure-Based prediction of the peptide sequence space recognized by natural and synthetic PDZ domains. J Mol Biol 402: 460–474.
46. Altman MD, Nalivaika EA, Prabu-Jeyabalan M, Schiffer CA, Tidor B (2008) Computational design and experimental study of tighter binding peptides to an inactivated mutant of HIV-1 protease. Proteins 70: 678–694.
47. Joachimiak LA, Kortemme T, Stoddard BL, Baker D (2006) Computational design of a new hydrogen bond network and at least a 300-fold specificity switch at a Protein-Protein interface. J Mol Biol 361: 195–208.
48. Reina J, Lacroix E, Hobson SD, Fernandez-Ballester G, Rybin V, et al. (2002) Computer-aided design of a PDZ domain to recognize new target sequences. Nat Struct Mol Biol 9: 621–627.
49. Reynolds KA, Hanes MS, Thomson JM, Antczak AJ, Berger JM, et al. (2008) Computational redesign of the SHV-1 beta-lactamase/beta-lactamase inhibitor protein interface. J Mol Biol 382: 1265–1275.
50. Gainza P, Roberts KE, Donald BR (2012) Protein design using continuous rotamers. PLoS Comput Biol 8: e1002335.
51. Goldstein R (1994) Efficient rotamer elimination applied to protein side-chains and related spin glasses. Biophys J 66: 1335–1340.
52. Piserchio A, Fellows A, Madden DR, Mierke DF (2005) Association of the cystic fibrosis transmembrane regulator with CAL: structural features and molecular dynamics. Biochemistry 44: 16158–16166.
53. Word JM, Lovell SC, Richardson JS, Richardson DC (1999) Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. J Mol Biol 285: 1735–1747.
54. Case DA, Cheatham TE, Darden T, Gohlke H, Luo R, et al. (2005) The amber biomolecular simulation programs. J Comput Chem 26: 1668–1688.
55. Word JM, Lovell SC, LaBean TH, Taylor HC, Zalis ME, et al. (1999) Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. J Mol Biol 285: 1711–1733.
56. Lazaridis T, Karplus M (1999) Effective energy function for proteins in solution. Proteins 35: 133–152.
57. Weiner SJ, Kollman PA, Nguyen DT, Case DA (1986) An all atom force field for simulations of proteins and nucleic acids. J Comput Chem 7: 230–252.
58. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, et al. (1983) CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 4: 187–217.
59. Cushing PR, Fellows A, Villone D, Boisgurin P, Madden DR (2008) The relative binding affinities of PDZ partners for CFTR: a biochemical basis for effcient endocytic recycling. Biochemistry 47: 10084–10098.
60. Vouilleme L, Cushing PR, Volkmer R, Madden DR, Boisguerin P (2010) Engineering peptide inhibitors to overcome PDZ binding promiscuity. Angew Chem Int Ed Engl 49: 9912–9916.
61. Dayhoff M, Schwartz R, Orcutt B (1978) A model of evolutionary change in proteins. In: Dayhoff M, ed. Atlas of Protein Sequence and Structure, vol. 5, suppl. 3. Washington, DC: Natl Biomed Res Found. pp 345–352.
62. Bruscia E, Sangiuolo F, Sinibaldi P, Goncz KK, Novelli G, et al. (2002) Isolation of CF cell lines corrected at DeltaF508-CFTR locus by SFHR-mediated targeting. Gene Ther 9: 683–685.
63. Li Y, Wang W, Parker W, Clancy JP (2006) Adenosine regulation of cystic fibrosis transmembrane conductance regulator through prostenoids in airway epithelia. Am J Respir Cell Mol Biol 34: 600–608.
64. Taddei A, Folli C, Zegarra-Moran O, Fanen P, Verkman AS, et al. (2004) Altered channel gating mechanism for CFTR inhibition by a high-affinity thiazolidinone blocker. FEBS Lett 558: 52–56.
65. Ma T, Thiagarajah JR, Yang H, Sonawane ND, Folli C, et al. (2002) Thiazolidinone CFTR inhibitor identified by high-throughput screening blocks cholera toxin-induced intestinal uid secretion. J Clin Invest 110: 1651–1658.
66. Saro D, Li T, Rupasinghe C, Paredes A, Caspers N, et al. (2007) A thermodynamic ligand binding study of the third PDZ domain (PDZ3) from the mammalian neuronal protein PSD-95. Biochemistry 46: 6340–6352.
67. Wiedemann U, Boisguerin P, Leben R, Leitner D, Krause G, et al. (2004) Quantification of PDZ domain specificity, prediction of ligand affinity and rational design of super-binding peptides. J Mol Biol 343: 703–718.
68. Gordon DB, Hom GK, Mayo SL, Pierce NA (2003) Exact rotamer optimization for protein design. J Comput Chem 24: 232–243.
69. Lippow SM, Wittrup KD, Tidor B (2007) Computational design of antibody-affinity improvement beyond in vivo maturation. Nat Biotech 25: 1171–1176.
70. Hom GK, Mayo SL (2006) A search algorithm for fixed-composition protein design. J Comput Chem 27: 375–378.
71. Sampson HM, Robert R, Liao J, Matthes E, Carlile GW, et al. (2011) Identification of a NBD1-Binding pharmacological chaperone that corrects the trafficking defect of F508del-CFTR. Chem Biol 18: 231–242.
72. Sheppard DN (2011) Cystic fibrosis: CFTR correctors to the rescue. Chem Biol 18: 145–147.

73. Gorczynski MJ, Grembecka J, Zhou Y, Kong Y, Roudaia L, et al. (2007) Allosteric inhibition of the protein-protein interaction between the leukemia-associated proteins runx1 and CBFbeta. Chem Biol 14: 1186–1197.

74. te Velthuis AJW, Sakalis PA, Fowler DA, Bagowski CP (2011) Genome-Wide analysis of PDZ domain binding reveals inherent functional overlap within the PDZ interaction network. PLoS ONE 6: e16047.