Local interactions that contribute minimal frustration determine foldability

Taisong Zou[1], Brian W. Woodrum[2*], Nicholas Halloran[2], Paul Campitelli[1], Andrey A. Bobkov[3], Giovanna Ghirlanda [2#] and S. Banu Ozkan[1#]

[1] Department of Physics and Center for Biological Physics, Arizona State University, Tempe, Arizona, 85287-1504

[2] School of Molecular Sciences, Arizona State University, Tempe, Arizona, 85287-1604

[3] Conrad Prebys Center for Chemical Genomics, SBP Medical Discovery Institute, La Jolla, CA

Corresponding Authors : Banu.Ozkan@asu.edu and Giovanna.Ghirlanda@asu.edu

ABSTRACT

Earlier experiments suggest that the evolutionary information (conservation and coevolution) encoded in protein sequences is necessary and sufficient to specify the fold of a protein family. However, there is no computational work to quantify the effect of such evolutionary information on the folding process. Here we explore the role of early folding steps for the sequences designed using coevolution and conservation through a combination of computational and experimental methods. We simulate a repertoire of native and designed WW domain sequences to analyze early local contact formation and find that the N-terminal beta-hairpin turn would not form correctly due to strong, non-native local contacts in unfoldable sequences. Through a maximum likelihood approach, we identify five local contacts that play a critical role in folding, suggesting that a small subset of amino acid pairs can be used to solve the "needle in the haystack" problem to design foldable sequences. Thus, using the contact probability of those five local contacts which form during the early stage of folding, we built a classification model that predicts the foldability of a WW sequence with 81% accuracy. This classification model is used to re-design WW domain sequences that cannot fold due to frustration and make them foldable by introducing a few mutations that lead to stabilization of these critical local contacts. The experimental analysis shows that a re-designed sequence folds and binds to polyproline peptides with similar affinity to those observed for native WW domains. Overall, our analysis shows that evolutionary designed sequences

should not only satisfy folding stability but also ensure a minimally frustrated folding landscape.

INTRODUCTION

Classic work by Nobel-laureate Christian Anfinsen showed that, for many proteins, the amino acid sequence alone is sufficient information to fold into a unique 3-D structure, the native structure. A brief combinatorial calculation of amino acid interactions shows that the information density of a protein sequence is vast. Specifying fold and function requires the impossible exploration of this enormous sequence space in totality. However, evolution, which is three billion years of Monte Carlo steps, has provided the data to dramatically reduce this search[1,2]. It is plausible that all the information required for specifying the fold and characteristic function of a protein may be sufficiently encoded in the small set of amino acid interactions revealed by coevolution and conservation analysis of a given fold. Using maximum entropy and/or Bayesian inference type methods on multiple sequence alignments of a given fold, one can incorporate important coevolved pairs as contacts, restraints in computer simulations [3–5]. This greatly reduces the search through sequence space which allows one to obtain the 3-D native structure of proteins, intermediate conformations, protein-ligand and protein-protein interactions [5–16].

The success of approaches utilizing evolutionary analysis also prompts the question of whether we can use coevolution and/or conservation (i.e. amino acid

preference of a position) information to design a foldable sequence. That is, can we measure and subsequently predict the foldability of a designed sequence based on evolutionary analysis?

Promising results were obtained previously for artificially designed sequences by using conservation and coevolution statistics based on the multiple sequence alignment (MSA) of 120 members of the natural WW domain family [17,18]. However, only about one third of the evolutionary designed sequences fold to the natural WW domain structure and display similar binding affinity and specificity; the remaining two thirds are unable to fold.

We hypothesize that the key in this discrepancy lies in the mechanism by which coevolved positions contribute to the folding energy, because only a small subset of coevolved contacts shapes the folding energy landscape. Identifying these contacts requires new methods that can isolate the set of coevolved positions in the sequence which encodes the depth and ruggedness of the folding funnel landscape [14,19]

In this work, we revisited these evolutionary designed WW sequences by exploring the emergence of interactions between the nearest positions in the sequence (i.e. local contacts), rather than focusing on the overall fold and its stability. Our results show that all foldable sequences form a nucleation site containing very strong native local contacts in the N-terminal β-turn, characteristic of WW domains. Our analysis also shows that, in unfoldable sequences, the N-terminal β-turn formation weakens due to formation of other

stable non-native contacts, leading to a frustration in the folding [20–22]. Using a maximum likelihood approach, we identify five local contacts that emerge at the beginning of the folding process critical for folding as they strongly contribute to the smoothness of the funneled landscape. By computing the probability of forming these five contacts, we successfully predict the foldability of a WW sequence. Building on the results of existing WW domain sequence studies, we develop a computational approach to design new foldable sequences by introducing mutations to unfolded sequences at positions forming these five crucial contacts. Several of these mutants were experimentally expressed; one sequence of particular note, CC16_N21, folds into the typical WW structure and displays thermal stability comparable to native sequences. Most importantly, CC16_N21 binds target polyproline peptides with affinity in the low micromolar range, akin to those observed for native WW domains. These results indicate that functional WW domains can be designed by optimizing contacts emerging earlier in the folding pathway rather than so-called native contacts.

METHODS

Computational Methods

Molecular Dynamics Simulations: We performed independent simulations of full-coverage 8-mer fragments (1st step of ZAM) of the sequences listed in Table SI to identify possible nucleation sites at the beginning of the folding process (further details provided in Supplementary Information).

Contact Probability Metric: Contact probability (CPROB) is the equilibrium probability of a given contact, calculated as the fraction of residue-to-residue alpha carbon separation distances less than 8Å. The CPROB of a contact *i* called x_i in a sequence is defined by the average CPROB over all the 8-mer fragments containing this contact. Thus each sequence has a vector of CPROBs over all possible N local contacts $\vec{x} = (x_1, x_2, \ldots, x_N)$, N being the total number of local contacts.

Classification Model: The foldability of each sequence in our dataset was known from earlier experiments and thus each sequence had a vector of CPROBs. Based on these data, we wished to train a probabilistic model to estimate the likelihood of a sequence being foldable versus unfoldable, given only the CPROBs observed in the 8-mer fragment simulations. This was a binary classification problem, where we had an unknown outcome z that could be either foldable (z = 1) or unfoldable (z = 0) and we wanted to calculate $(z = 1|\vec{x})$, the probability of the sequence being foldable given its CPROB vector (Pfold). Such a problem could be solved using a logistic regression model where the log odds (logit), a function of $P(z = 1|\vec{x})$, was assumed to be linearly related to $\vec{x}$:

$$\log \frac{P(z = 1|\vec{x})}{1 - P(z = 1|\vec{x})} = \alpha + \vec{\beta} \cdot \vec{x} \quad (1)$$

Solving for $P(z = 1|\vec{x})$ yields:

$$P(z=1|\vec{x}) = \frac{\exp\left(\alpha+\vec{\beta}\cdot\vec{x}\right)}{1+\exp\left(\alpha+\vec{\beta}\cdot\vec{x}\right)} \quad (2)$$

The linear coefficients α and $\vec{\beta}$ were estimated using a maximum likelihood estimation. The Wald statistics of β_i indicated the significance of the contact *i* (See Supplementary Information for details).

Design of Foldable WW Domain Sequences: We took an unfolded sequence as a scaffold and tried to maximize the expected Pfold $P(z=1|x_1, x_2 \cdots x_5)$ for the template by swapping its five crucial local contacts (ten residues) identified from the classification model with those of a foldable natural sequence. To achieve this, we enumerated all possible combinations of swaps (*i.e.*, swapping only one certain contact or two contacts, etc). The expected Pfold after swapping was calculated with eq. 2, where the CPROBs of the swapped contacts were represented by those from the foldable natural sequence and unswapped ones were kept as originally in the unfolded sequence. The hybrid sequence (a mixture of an unfolded template and amino acids from folded sequences) corresponding to the maximum expected Pfold would be further examined in ZAM simulation and experiment.

Experimental methods

The CC16_N21 designed WW peptide was ordered from Genscript in addition to 5 separate genes containing single point mutations from the N21 WW peptide. All genes were designed to be expressed as fusion proteins to the Maltose

Binding Protein (MBP) and were delivered in the pMAL-c5x vector for expression. For purification, each gene contained an N-terminal polyhistidine tag and the TEV cleavage site ENLYFQG. The DNA was transformed via heat shock into competent BL-21 cells and grown on Agar-LB plates containing ampicillin overnight at 37°C. Single colonies were used to inoculate 8 mL LB liquid cultures containing ampicillin and were grown overnight at 37°C shaking at 210 RPM. 1 mL of these cultures was transferred to a 2L flask containing 1L LB media with ampicillin for growth and expression. The rest of the cells were centrifuged down, and the plasmid DNA was extracted using Promega Wizard ® Plus SV Miniprep kits. Correct sequences were verified using ASU Sanger Sequencing. The 1L cultures were grown to an OD of 0.6 and protein expression was induced by addition of 1mM IPTG. Proteins were expressed for 3.5 hours at 37°C shaking at 210 RPM. Total protein yield for these conditions was roughly 20 mg/L, but induction at 0.8-0.9 OD increased the yield to 30-35 mg/L total protein without additional issues in the purification.

After Expression cells were pelleted out by centrifugation at 5,000 RPM for 20 minutes. 1L of cells were resuspended in 30mL pH 7.5 buffer containing 20mM Sodium Phosphate, 0.5M NaCl, and 40mM Imidazole. Cells were lysed by sonication, and then spun down at 5,000 RPM, 4°C for 30 minutes. The supernatant containing the cytoplasmic fraction was then flowed injected onto a 5mL Amersham Bioscience HisTrap column for purification. Bound protein was washed with 75mL resuspension buffer to remove unwanted proteins, and then eluted by 500mM in 7.5 buffer containing 20mM Sodium Phosphate, and 0.5M

NaCl. Eluted protein was recovered in 5mL fractions and quantified by UV-Vis spectroscopy using a Cary 50 Bio spectrophotometer.

Purified fusion proteins were desalted using GE Healthcare PD-10 desalting columns and subjected to buffer exchange into the original resuspension buffer for TEV cleavage and further purification compatibility. His-tagged TEV was added to the fusion proteins at a ratio of 1:10 TEV to fusion protein. After addition of 1mM DTT and 0.5mM EDTA, the reaction was left overnight at 4°C and peptides were further purified on the HisTrap column by collecting flow through.

MALDI-TOF was used to verify that each of the peptides was the correct mass, but all samples showed impurities in RP-HPLC. Peptides were further purified by RP-HPLC with a 250 x 10mm Phenomenex C18 Semi-prep column by gradient elution starting with .01% TFA in water to 95% acetonitrile with .01% TFA. Fractions were collected, frozen, and lyophilized to yield pure protein with an overall expression yield of about 2-3 mg/L.

Group I, N21, and CC16 peptides were synthesized on Wang resin using a CEM Liberty automated peptide synthesizer with FMOC protected amino acids. Using DMF as a solvent, 20% Piperidine with 0.1M HOBT was used to deprotect amino acids. Activation and coupling were completed using 0.5M HBTU and 2M DIEA in NMP respectively. Complete cleavage was accomplished by shaking for 2 hours using a cleavage cocktail containing 90% TFA, 5% Thioanisole, 3% DODT,

and 2% Anisole. After cleavage, peptides were precipitated with cold ether, pelleted by centrifugation, and washed four times. Purification was done by RP-HPLC using the same protocol for the expressed WW peptides above. Correct proteins were verified by MALDI-TOF after purification.

Protein stability and folding was analyzed using a JASCO J-815 CD Spectrophotometer. Full scans were measured from 260nm-200nm at 5°C using a 1cm quartz cuvette, which showed the distinctive WW domain signal at around 229nm. T_{melt} for all the WW peptides were calculated by monitoring ellipticity at 229nm while increasing temperature from 5°C to 90°C at 0.5°C/min. ITC titrations were performed using ITC200 (GE Lifesciences) instrument at 4°C. One of 1 µl and 18 of 2 µl injections of 5.1 mM Group I peptide were made into the cell containing either 0.102 mM native N21 or 0.097 mM CC16_N21. The ITC data were fitted using MIcrocal Origin software provided by the ITC manufacturer.

RESULTS

Crucial local contacts highly impact foldability

Most earlier work on designing foldable protein sequences focused on optimizing structural stability to ensure foldability. However, it has been shown that it is also possible to design foldable artificial sequences by inferring coevolved positions. Designing artificial sequences by inferring coevolution and conservation from multiple sequences has been applied first to WW domains[17,23]. Two libraries of artificial sequences were constructed using computational algorithms: (i) site-

independent conservation (IC) sequences which only preserve the amino acid composition (conservation) at each single site but diminish the pairwise coevolution between sites and (ii) coupled-conservation (CC) sequences which maintain both the pattern of conservation and pairwise coevolution information. Additionally, native (N) sequences have been used for comparison throughout the study. While constructed IC sequences and CC sequences shared similar sequence identities to those of native sequences, only 1/3 of these constructed sequences were able to fold.

To explore how such evolutionary information specifies a protein fold, we analyzed previously designed native-like sequences of the WW domain. Particularly we focused on initial local contact formation patterns of the foldable and unfoldable sequences which follow distinct patterns. If these patterns are able to distinguish foldable from unfoldable sequences, the prediction of foldability may be possible through emergence of these local contacts. Supporting this approach, our earlier studies on protein folding show that local contacts are critical for foldability, and proteins initiate a fold by forming independent local fragments on the shortest time scales (i.e. zipping steps). [24,25]

For each sequence, we simulated short 8-mer fragments using replica exchange molecular dynamics (REMD) with the AMBER ff96 force field [26] and the Generalized Born (GB) implicit solvent model [27] and computed the equilibrium contact probability (CPROB) between any two positions (*i.e.*, the probability that a contact is formed) in the equilibrium simulations. If a contact is included in

many fragments, the CPROB of the contact is computed by averaging CPROB over all the 8-mer fragments containing this contact (i.e. the contact between 3rd and 7th residue position is computed using fragments 1-8, 2-9, 3-10). For each sequence, we have a CPROB vector over all possible N local contacts where $\vec{x} = (x_1, x_2, \ldots, x_N)$. (here, N=115, based on 28 positions that do not have gaps in MSA, see supporting information for details). We computed a total of 89 CPROB vectors arising from 40 foldable and 49 unfoldable sequences (See Supplementary Tab. S1). To compare the foldable sequences with unfoldable sequences, we also calculated the maximum likelihood CPROB (MLCPROB), obtained from the histogram of CPROB (Supplementary Fig. S1). We used a normal Kernel Density Estimate to smooth the histograms which removes the dependence on the bin starting points and better reflects the underlying data.

We then constructed MLCPROB maps for foldable and unfoldable sequences separately based on 8-mer fragment simulations. On the map, each rectangle represents a local contact colored by MLCRPOB values ranging from 0 (blue) to 1 (red). We found that native foldable sequences give rise to strong local interactions in the turn segment of the N-terminal hairpin based on high local contact probabilities observed in this region (Fig. 1(A)). On the contrary, unfoldable sequences give rise to weak local interactions in the N-terminal hairpin (Fig. 1(B)). Differences became further pronounced when these small 8-mer structures were grown into larger 16-mer fragments, clearly delineating two behaviors: (i) experimentally foldable sequences display the emergence of contacts corresponding to the N-terminal hairpin, measured with a high CPROB

score and (ii) experimentally unfoldable sequences display very low CPROB near N-terminal hairpin indicating weak interactions. In contrast, strong non-native interactions are observed in another region, possibly leading to a frustration in the landscape (i.e. decreasing the smoothness) (Fig.1 (C-D)). Indeed, the average contact map of all foldable sequences shows the same trend of formation of N-terminal hairpin contacts (Supplemental Fig S2). However, the local native contacts of the N-terminal hairpin turn are notably missing the in the contact map averaged over all unfoldable sequences. Overall, the emergence of strong native contacts that favors hairpin formation leads to correct folding in all foldable designed sequences. On the other hand, the contact probability of these contacts are rather weak in unfoldable sequences, due to emergences of non-native contacts and misfolding [28].

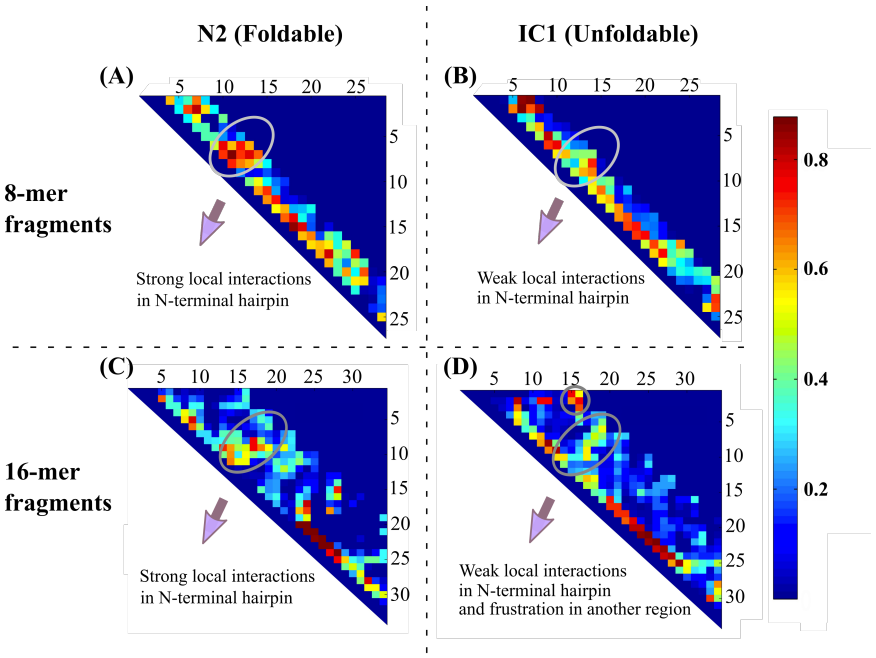


Figure 1. Average contact probability (obtained by MLCPROB) maps from 8-mer fragment simulations for **(A)** foldable native sequence N2 and **(B)** unfoldable sequence IC1. Foldable sequences show trends of forming local contacts around the turn of the N-terminal hairpin. Contact probability (CPROB) maps for larger 16-mer fragment simulations for **(C)** a representative foldable sequence N2 shows the same trend of strong formation of native N-terminal hairpin turn contacts in a foldable sequence, whereas **(D)** shows that an unfoldable sequence IC1 exhibits non-native interactions in other regions which weaken the formation of this N-terminal hairpin turn.

Contact probabilities of local interactions can predict foldability

Our observation that foldable and unfoldable sequences show different patterns of emergence of local contacts at the early stage of protein folding suggests that the contact probabilities of these local contacts (i.e. contact probability vector of 115 local contacts) can be utilized to predict whether a sequence is foldable or not. A further reduction to a minimum set of specific local contacts could prove much more relevant for the initiation of folding (nucleation sites) than others by making foldability predictions which only implement the contact probability of this minimum subset, particularly for evolutionary designed sequences. To verify this, we built a classification model where the probability of a sequence being foldable given by its contact probability (CPROB) vector expressed as a function of $\vec{x}$ and

solved using maximum likelihood estimation (see Methods and Supplementary Information for details).

Using this classification analysis, we found that a minimum set of five out of the 115 local contact elements of the CPROB vector $\vec{x}$ are enough to differentiate the foldability of sequences in the dataset with high accuracy. These five crucial contacts are listed in Table I. We computed the CPROB of these five local contacts from simulations employing our zipping and assembly methodology (ZAM)[24,29]. The CPROB data was then used to predict the foldability of WW sequences with an average true prediction rate of 80.9%, (Table I) where the sequences are classified to be foldable or unfoldable if the conditional probability of foldability (Pfold) $P(z = 1|x_1, x_2 \cdots x_5) > 0.5$ or < 0.5, respectively. This model also shows excellent statistical significance compared with random models using any five of the 115 contacts, with a high true prediction rate (Fig. S2 (A)) and low deviance (Fig. S2 (B)). Mapping those contacts onto a crystallographic structure of WW domain, we found that four of them are located in or around the N-terminal hairpin, which has been shown to be critical in the folding process in many experiments and simulations [24,30–34]. Interestingly, one of these contacts is a non-native interaction and thus has a negative contribution to foldability score (Fig. 2).

Table I. Statistics of 5 crucial local contacts found in the classification model. (S.E. is short for standard error associated with β coefficients in the classifier,

and Wald Statistics and p-value quantify the significance of parameters in our classifier)

Crucial Contacts	Coefficient	S.E.	Wald Statistics	*p*-value
(2,7)	3.426	1.434	5.713	0.017
(4,7)	-6.278	1.770	12.581	0.000
(10,13)	-3.330	1.234	7.282	0.007
(11,16)	6.685	1.995	11.233	0.001
(25,28)	5.554	1.843	9.086	0.003

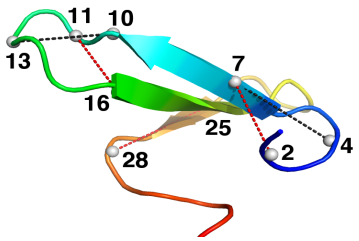


Figure 2. Mapping five crucial local contacts onto the 3-D WW domain structure. The contact formation probabilities (CPROB) of these five contacts enables us to predict foldability of designed WW sequences with a TP rate of~80.9 Four local contacts are located around the N-terminal hairpin of WW domain and the contact between 25-28 is a negative control.

Given that the formation of N-terminal hairpin is a critical step of folding, and the importance of five local contacts found from statistical analysis on simulation data, we next wanted to determine whether the stabilization of those crucial local contacts could assist the formation of the N-terminal hairpin and subsequently promote folding. To test this idea, we artificially constrained two crucial local contacts ((10, 13) and (11, 16)) at the N-terminal hairpin in the simulation for several unfolded sequences. These non-native constraints did indeed increase the probability of forming the N-terminal hairpin, leading to the correct fold when growing the chain from 8-mer fragments in ZAM simulations [24,29] (Fig. 3). In ZAM simulations, each 8-mer fragment undergoes a 5 ns per replica REMD [35] starting from a fully extended conformation. We analyzed the results by using weighted histogram analysis and identified the fragments which form stable

hydrophobic contacts with well-formed turns or helical shapes, as determined from the potential of mean force (PMF). We then loosely enforced those contacts with added restraints and grew 8-mer fragments by adding more residues in extended form. New REMD simulations were then performed on those larger fragments. A new PMF analysis is performed to see whether new hydrophobic contacts are formed, till full sequence is obtained.

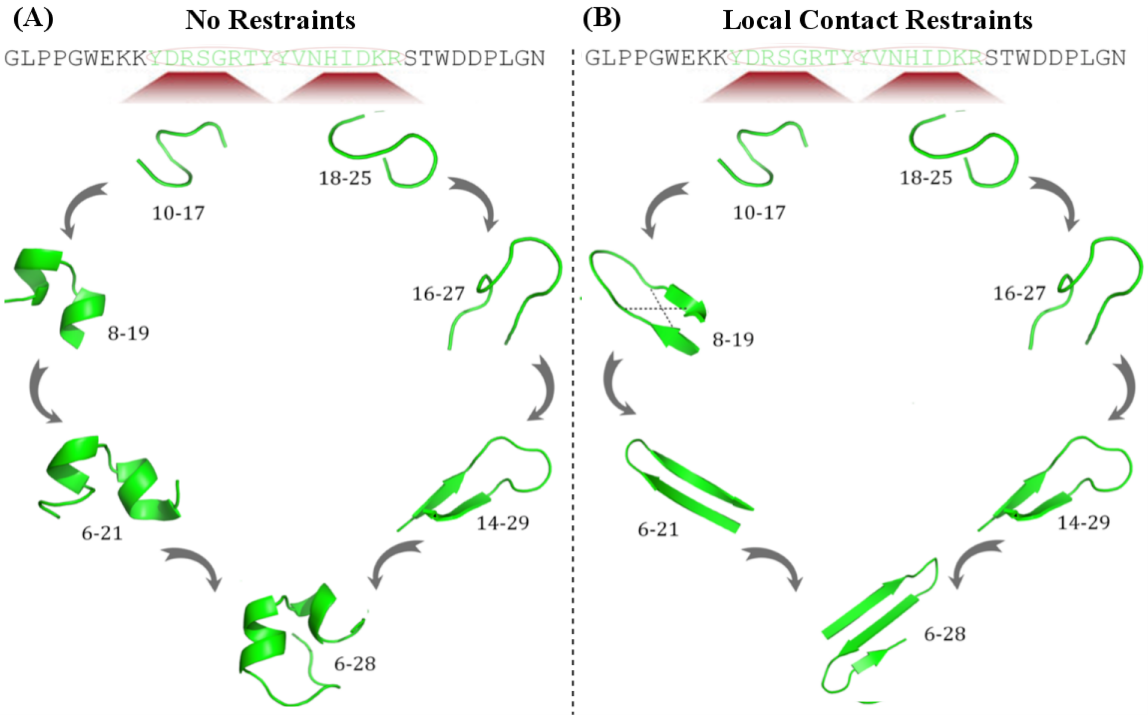


Figure 3. (A) The folding pathway of an unfolded WW sequence (CC36) using ZAM. CC36 turns out misfolded in the simulation. **(B)** Adding constraints to the crucial local contacts helps form the N-terminal hairpin correctly and make this unfolded sequence foldable.

A new design approach by utilizing the crucial contacts of foldability

Overall, predicting the foldability of a sequence based on local contact probabilities suggests that entropically low-cost local contacts contribute to folding and are likely coevolved due to their importance in providing smoothness to the folding landscape. Yet it may be difficult to get sufficiently strong signaling from coevolutionary analysis to identify “the direct local contacts” that contribute to the smoothness from indirect local contacts due to noise inherent in MSAs. Our maximum entropy approach through computing local contact formation can be coupled with coevolutionary inference analysis in this regard to redesign unfoldable sequences and make them foldable.

Our protocol differs fundamentally from the conventional computational protein design approach in which energy or stability optimization is used to search for ideal amino acid sequences given a fixed backbone topology (structure). Instead we identify mutations to the amino acid pairs at the positions of the five crucial local contacts (10 total residues) to maximize expected folding probability, Pfold $P(z = 1|x_1, x_2 \cdots x_5)$.

Here, we took each unfolded sequence as a scaffold and attempted to maximize the expected Pfold for the template by swapping its five local contacts (or ten residues) with those of a foldable natural sequence. To achieve this, we enumerated over all possible combinations of swaps (i.e., swapping only one certain contact or two contacts, etc.). The expected Pfold after swapping was calculated with the obtained classifier where the CPROBs of the swapped contacts were represented by those from the foldable natural sequence and unswapped ones were kept as originally present in the unfolded sequence. Then

these hybrid sequences (now a variant of CC unfoldable sequences with amino acids substitutions from foldable native sequences) corresponding to the maximum expected Pfold 0.9 or greater were further examined in ZAM simulations. Using our ZAM protocol, the 8-mer fragments of these hybrid sequences were simulated to obtain the CBPROB values at the critical positions and their Pfold values were then re-estimated (Fig. S3). With this approach we have generated 227 hybrid sequences. We then selected those fragments with high Pfold values (i.e. folding probabilities of at least 0.7) to grow to larger 16-mer fragments. At the 16-mer fragment step, we filtered out those failing to form N-terminal hairpins and grew the rest to the full sequences. Finally, we chose the sequences with correctly folded WW structures (i.e. those having backbone RMSD < 3 Å) as the foldable sequence candidates, which was subjected to experimental verification. (Fig. 4)

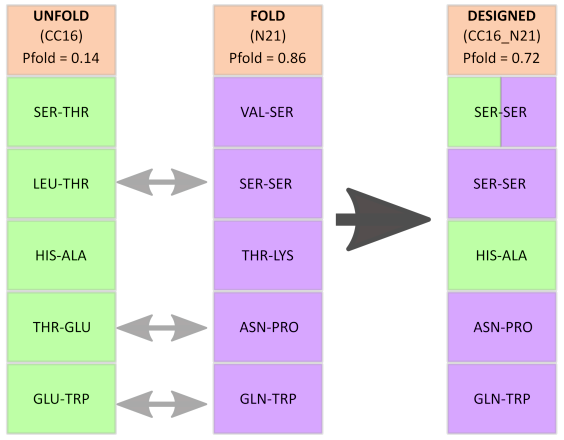


Figure 4. The designed sequence CC16_N21 was generated based on an unfolded scaffold CC16 and a folded sequence N21. One of the five critical

contacts remains the same as in CC16 (green) and three contacts are chosen to swap (purple). The threonine substituted to serine in the second contact of the designed sequence also resulted in a hybrid serine-serine contact in the designed sequence (green/purple) as a result of this residue (residue 7) taking part in two crucial contacts. The unfolded, folded and designed sequences are displayed below for visual inspection.

Redesigned CC16_N21 based on crucial local contacts folds and function as natural WW domains

The sequence with the highest foldability based on computational analysis, CC16-N21, was selected for biophysical characterization. We compared the secondary structure and stability to thermal denaturation of CC16_N21 to that of the parent unfoldable sequence CC16 and that of the donor native WW domain sequence, N21, by circular dichroism (CD). The secondary structure of the designed CC16_N21 protein resembles that of N21 and CC16, as shown by similar features in the CD spectrum including a maximum in the CD spectrum at approximately 227 nm (Fig. 5 (A)), a signature for WW domains; as expected, CC16 lacks these features and appears unfolded.

Although the native contacts introduced in CC16_N21 are sufficient to restore the signatures of a WW fold, the designed protein is less stable to thermal denaturation than its native counterpart, N21. Thermal denaturation curves were obtained by monitoring the loss of CD signal at 227 nm as a function of temperature in the 4˚C to 90˚C range (Fig. 5 (B)), and yielded an apparent T_m of

46.7˚C for N21, and of 22.4 ˚C for CC16_N21; CC16 shows no transition, due to its unfolded structure.

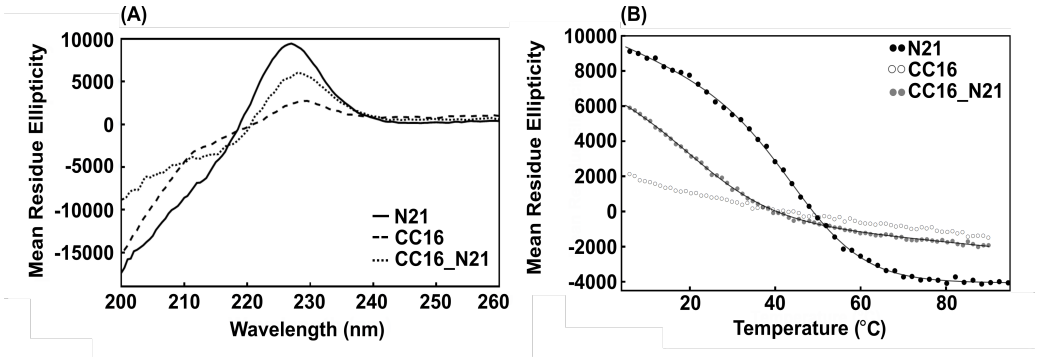
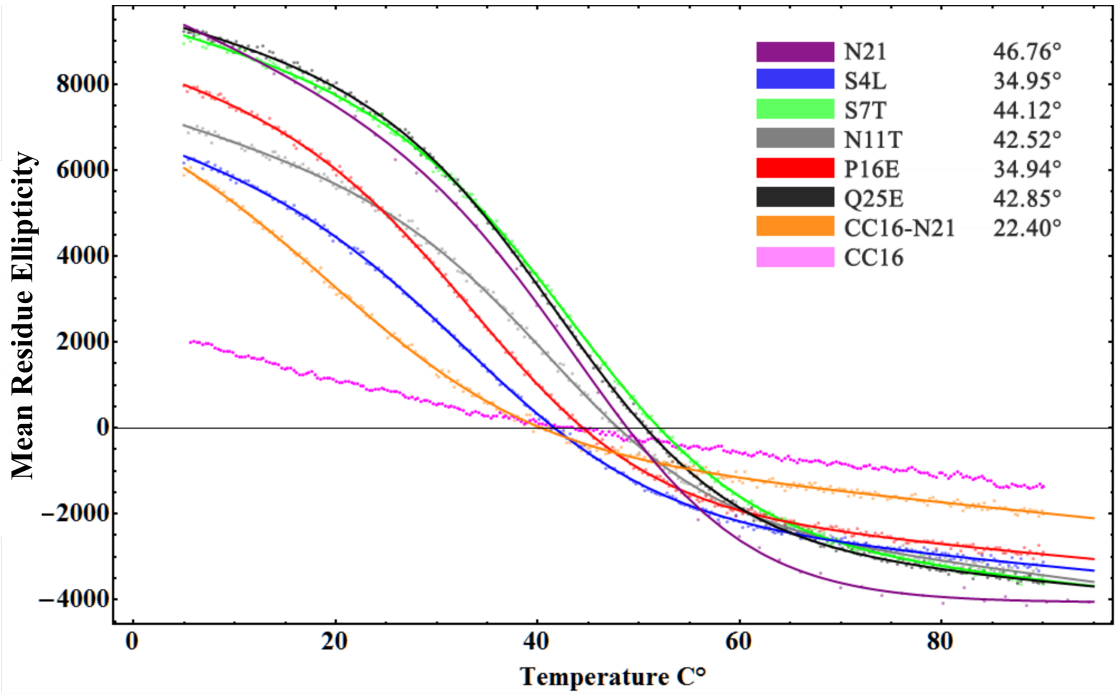


Figure 5. Structure and stability of WW domains. **(A)** CD spectra of N21 (solid black line), CC16 (dashed black line), CC16_N21 (dotted black line) collected at 4˚C. Both spectra display a peak at 227 nm typical of folded WW domain. **(B)** Thermal denaturation curves of N21 (black circles), CC16 (white circles), and CC16_N21 (grey circles), and CC16, white circles. Conditions: phosphate buffer 20 mM, pH 7.

Compared to the parent sequence CC16, which is unfolded [17,18], our re-designed version of CC16, (CC16_N21) folds into the native WW domain fold, as shown by its CD spectrum at 4 ˚C. When compared to the native N21 sequence, CC16_N21 displays lower stability as shown by the temperature denaturation curves. However, the observed apparent T_m of 22.4 ˚C is well within the range observed for natural WW domains [36,37].

An alternative explanation for the observed rescue of foldability in CC16 is that the five amino acids in N21 at these positions are important for stability of the WW fold. To deconvolute the contribution of each mutation to the stability of the WW fold, we introduced each of the five amino acids that differentiate CC16 from N21 individually on the background of N21 (Fig 6). We found that three mutations, S7T, N11T, and Q25E, are relatively conservative and have minimal effect on stability; two mutations (S4L and P16E) that replace a hydrophilic residue with a polar one, and a proline with glutamic acid respectively, destabilize the fold to some extent. These results indicate that most of the mutations inserted into CC16 would have little or no effect on stability.



N21	GSVESDWSVHTNEKGTPYYHNRVTKQTSWIKPDVL
CC16-N21	GGSK**S**GW**S**EYH**N**DAGT**P**YYYDR**Q**TGQSTWEKPEDF
CC16	GGSKLGWTEYHTDAGTEYYYDRQTGESTWEKPEDF
N21-S4L	GSVE**L**DWSVHTNEKGTPYYHNRVTKQTSWIKPDVL
N21-S7T	GSVESDW**T**VHTNEKGTPYYHNRVTKQTSWIKPDVL
N21-N11T	GSVESDWSVHT**T**EKGTPYYHNRVTKQTSWIKPDVL
N21-P16E	GSVESDWSVHTNEKGT**E**YYHNRVTKQTSWIKPDVL
N21-Q25E	GSVESDWSVHTNEKGTPYYHNRVTK**E**TSWIKPDVL

Figure 6. CD monitored thermal denaturation curves of the parent proteins and of N21 mutants. Conditions: phosphate buffer 20 mM, pH 7; protein concentrations were: N11T 10.0 mM, S7T 5.75 mM, Q25E 14.3 mM, P16E 9.28 mM, S4l 18.58 mM, CC16-N21 16.35 mM, CC16 12.0 mM, N21 7.55 mM.

We next investigated whether the designed CC16_N21 sequence conserves the function of native WW domains. These domains typically recognize proline-rich peptides belonging to four distinct classes with micromolar affinity and some

degree of in-class specificity. Given that the vast majority of WW domains display at least some degree of binding affinity for Class I peptides, we assessed the fitness of both CC16_N21 and N21 by evaluating binding to a model Class I peptide, EYPPYPPPPYPSG, using isothermal titration calorimetry (ITC) at 4˚C (Fig. 7). We found that the native N21 did not bind Group I peptide up to high micromolar concentrations, despite previous literature data indicating weak binding by oriented peptide array [17,36,38]. In contrast, fitting of the titration curve for our designed sequence CC16_N21 resulted in a K_d = 71.0 μM ± 4.7 μM, comparable to those recorded for native WW domains, typically in the 10-300 μM range [17,36,38].

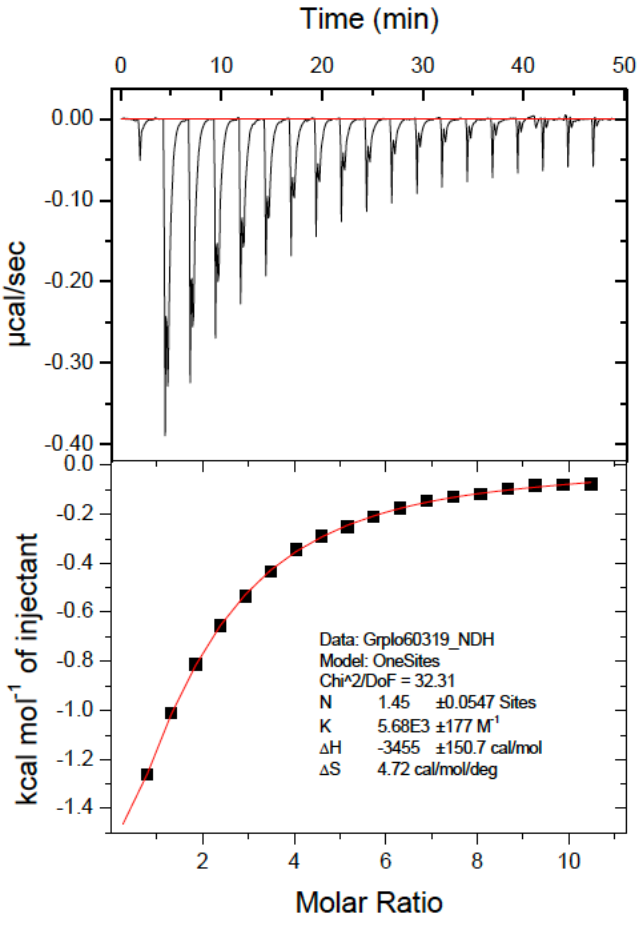


Figure 7. Isothermal titration calorimetry curve of Group I model peptide EYPPYPPPPYPSG into CC16_N21 100 μM, phosphate buffer 20 mM pH 7. Fitting of the data suggests one binding site and duplicate titrations give 73.5 μM ± 4.4 μM and 68.5 μM ± 5.0 μM.

DISCUSSION

We studied the interactions between near neighbor residues in the sequence (i.e. local contacts) in early steps of the folding process using a repertoire of evolutionary designed WW domain sequences, and explored the role of local contact formation pattern in predicting their foldability. Particularly, we investigated why only one third of the sequences designed based on coevolution and conservation fold, even while sharing high sequence similarity to those of native foldable sequences. When we analyzed the local contact emergence between foldable and unfoldable sequences, we observed that minimum frustration plays a crucial role for a given sequence's ability to fold [39]. The foldable sequences follow the experimentally observed folding, where strong local native contacts emerge around the N-terminal hairpin turn. On the other hand, the native local contact formation is weakened in unfoldable sequences due to non-native local interactions, suggesting the frustration in the landscape prevents folding.

Recent folding studies of resurrected ancestral RNaseH proteins have shown that the main folding intermediate is conserved over 3 Billion years of evolution, and folding pathways to reach the folding intermediate are modulated through the formation local contacts, particularly helical propensities [40,41]. Similarly, the folding kinetics of ancestral thioredoxins [39] along with the extant homologs suggested that evolution utilizes minimum frustration for folding. Overall, these studies are in agreement with our finding that there exist a minimum set of coevolved positions critical for early steps of folding which may play role for the smoothness of the potential energy landscape. Thus, when evolutionary inference is used to design novel sequences, conventional computational design approaches primarily depending on protein stability may fail to predict foldability and it should be complemented with approaches like ours that also incorporate folding kinetics.

We further tested whether conventional computational folding stability analyses could distinguish between the foldable and unfoldable sequences of WW domains. For this analysis, we first built model structures of these sequences using Modeller [42]. We then computed the folding stability of these models by commonly used computational folding stability methods such as FoldX [43] and DFIRE [44]. As expected, the folding stability distributions of foldable and unfoldable sequences yielded very similar trends (Fig. S5). We also obtained a classification model using the FoldX and DFIRE folding energy scores and obtained ROC curves in predicting foldability of sequences (Fig. S6 (A-B)). We found that both bioinformatics-based folding energy scores fails to predict

foldability of a given sequencing, yielding AUC around 0.52 (which is as good as random guessing). Likewise, other bioinformatics-based servers that implement biophysical properties of natively unfolded versus folded proteins such as FoldIndex [45] also yielded similar low prediction accuracy with an AUC of 0.63 (Fig. S6 (C)).

Besides bioinformatics-based approaches, we also performed 150 ns all-atom explicit water molecular dynamics simulation starting from the modeled structures for only three sequences, N21, CC16 and CC16_N21. Both unfoldable sequence CC16 and our re-designed CC16_N21 with an additional 4 mutations are similar to the foldable native sequence N21 with sequence similarities of 50 and 58 %, respectively. The RMSD plots of CC16, N21 and CC16_N21 (Fig. S7) show that they all exhibit fairly similar stabilities. Interestingly, the unfoldable sequence CC16 was shown to be the most stable over the simulation run. We also obtained energy landscapes of these three modeled structures as a function of native heavy-atom contacts and beta sheet secondary structure fractions as used previously to check stability [46] which all share similar shapes, again suggesting that the modeled folded structures exhibit similar stability. Surprisingly, the unfoldable CC16 sequence is slightly more stable than the foldable native N21 and designed CC16_N21 sequences. Overall these analyses suggest that all-atom molecular forcefields also fails to discriminate a foldable evolutionary designed sequence from that of an unfoldable sequence of a given fold if one starts with homology-modeled folded structures.

On the other hand, our use of local contact formation patterns, evaluated during the early stages of the folding process, can differentiate the foldable versus unfoldable sequences. Based on the contact probability of five local contacts using 8-mer fragment simulations, we built a classification model which could predict the foldability of WW domain sequences with high accuracy of 0.82. Enforcing the formation of certain local contacts in the WW domain also helps to avoid misfolding and leads to correct WW domain structures. Moreover, altering the contact probability formation of these five crucial contacts through amino acid substitutions, we have shown that it is possible to make frustrated unfolded sequences such as CC16 foldable.

This then raises the question as to why other designed sequences were frustrated and did not properly fold into a WW domain, despite the fact that they also contained the same coevolution (pairwise evolutionary coupling) and conservation (single site amino-acid frequency). One possible explanation is that discerning directly coupled coevolved contacts from indirectly coupled ones becomes challenging for local contacts as the data contains much more noise. While the coevolution information can be used to construct sequences with high stability once folded, the additional noise makes it extremely difficult to identify other obfuscating traps such as kinetic barriers which can exist at early stages in the folding process. Our approach allows for a functional and mechanistic analysis of early-folding contact formation which can provide additional information to highly complement evolutionary designed methods to create foldable units. Simply put, we enforce formation of the local contacts crucial for

early steps of folding. Then the designed sequences sharing the same coevolution and conservation of native sequences that cannot fold due to frustration become foldable, as these local contacts help smooth out the landscape (Fig. 8).

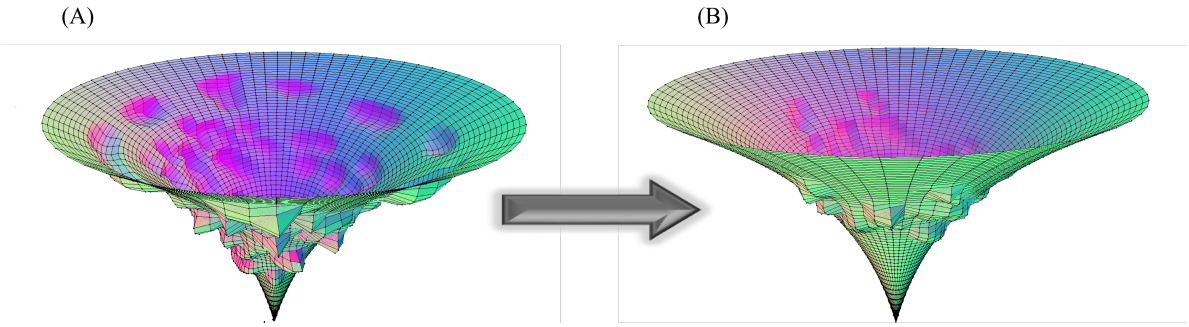


Figure 8. A cartoon model of folding energy funnel landscape. Here, jagged purple regions indicate local folding energetic minima. **(A)** The evolutionary designed WW sequences that cannot fold due to frustration in early states of folding with a frustrated landscape as compared to **(B)** our re-designed unfoldable sequences based on five local contacts that prevents early non-native contact formation, thus smoothing out the funnel.

SUPPORTING INFORMATION

Additional supplementary information is available which contains extended descriptions of computational and experimental methodologies as well as additional computational and experimental analysis and associated figures.

ACKNOWLEDGEMENTS

Support from NSF MCB Award 1715591 is gratefully acknowledged by SBO and GG. We also thank Andrei Bobkov, Protein Analysis Core, Sanford Burnham Prebys Medical Discovery Institute for his assistance in obtaining the binding affinity analysis of our designed WW domain. We also acknowledge computing time from Arizona State University Research Computing.

References

References

(1) Modi, T.; Campitelli, P.; Kazan, I. C.; Ozkan, S. B. Protein folding stability and binding interactions through the lens of evolution: a dynamical perspective. *Current opinion in structural biology* **2020**, *66*, 207–215. DOI: 10.1016/j.sbi.2020.11.007. Published Online: Dec. 31, 2020.

(2) Campitelli, P.; Modi, T.; Kumar, S.; Ozkan, S. B. The Role of Conformational Dynamics and Allostery in Modulating Protein Evolution. *Annual review of biophysics* **2020**, *49*, 267–288. DOI: 10.1146/annurev-biophys-052118-115517. Published Online: Feb. 19, 2020.

(3) Jana, B.; Morcos, F.; Onuchic, J. N. From structure to function: the convergence of structure based models and co-evolutionary information. *Physical chemistry chemical physics : PCCP* **2014**, *16* (14), 6496–6507. DOI: 10.1039/c3cp55275f. Published Online: Mar. 7, 2014.

(4) Juan, D. de; Pazos, F.; Valencia, A. Emerging methods in protein co-evolution. *Nature reviews. Genetics* **2013**, *14* (4), 249–261. DOI: 10.1038/nrg3414. Published Online: Mar. 5, 2013.

(5) Hopf, T. A.; Schärfe, C. P. I.; Rodrigues, J. P. G. L. M.; Green, A. G.; Kohlbacher, O.; Sander, C.; Bonvin, A. M. J. J.; Marks, D. S. Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife* **2014**, *3.* DOI: 10.7554/eLife.03430.

(6) Bai, F.; Morcos, F.; Cheng, R. R.; Jiang, H.; Onuchic, J. N. Elucidating the druggable interface of protein-protein interactions using fragment docking and coevolutionary analysis. *Proceedings of the National Academy of Sciences of the United States of America* **2016**, *113* (50), E8051-E8058. DOI: 10.1073/pnas.1615932113.

(7) dos Santos, R. N.; Morcos, F.; Jana, B.; Andricopulo, A. D.; Onuchic, J. N. Dimeric interactions and complex formation using direct coevolutionary couplings. *Scientific reports* **2015**, *5*, 13652. DOI: 10.1038/srep13652.

(8) Hayat, S.; Sander, C.; Marks, D. S.; Elofsson, A. All-atom 3D structure prediction of transmembrane β-barrel proteins from sequences. *Proceedings of the National Academy of Sciences of the United States of America* **2015**, *112* (17), 5413–5418. DOI: 10.1073/pnas.1419956112.

(9) Hopf, T. A.; Ingraham, J. B.; Poelwijk, F. J.; Schärfe, C. P. I.; Springer, M.; Sander, C.; Marks, D. S. Mutation effects predicted from sequence co-variation. *Nature biotechnology* **2017**, *35* (2), 128–135. DOI: 10.1038/nbt.3769.

(10) Hopf, T. A.; Morinaga, S.; Ihara, S.; Touhara, K.; Marks, D. S.; Benton, R. Amino acid coevolution reveals three-dimensional structure and functional domains of insect odorant receptors. *Nature communications* **2015**, *6*, 6077. DOI: 10.1038/ncomms7077.

(11) Marks, D. S.; Hopf, T. A.; Sander, C. Protein structure prediction from sequence variation. *Nature biotechnology* **2012**, *30* (11), 1072–1080. DOI: 10.1038/nbt.2419.

(12) Morcos, F.; Hwa, T.; Onuchic, J. N.; Weigt, M. Direct coupling analysis for protein contact prediction. *Methods in molecular biology (Clifton, N.J.)* **2014**, *1137*, 55–70. DOI: 10.1007/978-1-4939-0366-5_5.

(13) Morcos, F.; Pagnani, A.; Lunt, B.; Bertolino, A.; Marks, D. S.; Sander, C.; Zecchina, R.; Onuchic, J. N.; Hwa, T.; Weigt, M. Direct-coupling analysis of residue coevolution

captures native contacts across many protein families. *Proceedings of the National Academy of Sciences of the United States of America* **2011**, *108* (49), E1293-301. DOI: 10.1073/pnas.1111471108.

(14) Noel, J. K.; Morcos, F.; Onuchic, J. N. Sequence co-evolutionary information is a natural partner to minimally-frustrated models of biomolecular dynamics. *F1000Research* **2016**, *5.* DOI: 10.12688/f1000research.7186.1.

(15) Sułkowska, J. I.; Morcos, F.; Weigt, M.; Hwa, T.; Onuchic, J. N. Genomics-aided structure prediction. *Proceedings of the National Academy of Sciences of the United States of America* **2012**, *109* (26), 10340–10345. DOI: 10.1073/pnas.1207864109.

(16) Tang, Y.; Huang, Y. J.; Hopf, T. A.; Sander, C.; Marks, D. S.; Montelione, G. T. Protein structure determination by combining sparse NMR data with evolutionary couplings. *Nature methods* **2015**, *12* (8), 751–754. DOI: 10.1038/nmeth.3455.

(17) Russ, W. P.; Lowery, D. M.; Mishra, P.; Yaffe, M. B.; Ranganathan, R. Natural-like function in artificial WW domains. *Nature* **2005**, *437* (7058), 579–583. DOI: 10.1038/nature03990.

(18) Socolich, M.; Lockless, S. W.; Russ, W. P.; Lee, H.; Gardner, K. H.; Ranganathan, R. Evolutionary information for specifying a protein fold. *Nature* **2005**, *437* (7058), 512–518. DOI: 10.1038/nature03991.

(19) Morcos, F.; Schafer, N. P.; Cheng, R. R.; Onuchic, J. N.; Wolynes, P. G. Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. *Proceedings of the National Academy of Sciences of the United States of America* **2014**, *111* (34), 12408–12413. DOI: 10.1073/pnas.1413575111.

(20) Fisher, K. M.; Haglund, E.; Noel, J. K.; Hailey, K. L.; Onuchic, J. N.; Jennings, P. A. Geometrical Frustration in Interleukin-33 Decouples the Dynamics of the Functional Element from the Folding Transition State Ensemble. *PloS one* **2015**, *10* (12), e0144067. DOI: 10.1371/journal.pone.0144067.

(21) Nymeyer, H.; Socci, N. D.; Onuchic, J. N. Landscape approaches for determining the ensemble of folding transition states:Success and failure hinge on the degree of frustration. *Proceedings of the National Academy of Sciences of the United States of America* **2000**, *97* (2), 634–639.

(22) Oliveira, L. C.; Schug, A.; Onuchic, J. N. Geometrical features of the protein folding mechanism are a robust property of the energy landscape:A detailed investigation of several reduced models. *The journal of physical chemistry. B* **2008**, *112* (19), 6131–6136. DOI: 10.1021/jp0769835.

(23) Tian, P.; Louis, J. M.; Baber, J. L.; Aniana, A.; Best, R. B. Co-Evolutionary Fitness Landscapes for Sequence Design. *Angewandte Chemie (International ed. in English)* **2018**, *57* (20), 5674–5678. DOI: 10.1002/anie.201713220.

(24) Ozkan, S. B.; Wu, G. A.; Chodera, J. D.; Dill, K. A. Protein folding by zipping and assembly. *Proceedings of the National Academy of Sciences of the United States of America* **2007**, *104* (29), 11987–11992. DOI: 10.1073/pnas.0703700104.

(25) Zou, T.; Ozkan, S. B. Local and non-local native topologies reveal the underlying folding landscape of proteins. *Physical biology* **2011**, *8* (6), 66011. DOI: 10.1088/1478-3975/8/6/066011.

(26) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E.; DeBolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications* **1995**, *91* (1-3), 1–41. DOI: 10.1016/0010-4655(95)00041-D.

(27) Tsui, V.; Case, D. A. Molecular Dynamics Simulations of Nucleic Acids with a Generalized Born Solvation Model. *J. Am. Chem. Soc.* **2000**, *122* (11), 2489–2498. DOI: 10.1021/ja9939385.

(28) Ghosh, K.; Ozkan, S. B.; Dill, K. A. The ultimate speed limit to protein folding is conformational searching. *J. Am. Chem. Soc.* **2007**, *129* (39), 11920–11927. DOI: 10.1021/ja066785b.

(29) Shell, M. S.; Ozkan, S. B.; Voelz, V.; Wu, G. A.; Dill, K. A. Blind test of physics-based prediction of protein structures. *Biophysical journal* **2009**, *96* (3), 917–924. DOI: 10.1016/j.bpj.2008.11.009.

(30) A Beccara, S.; Škrbić, T.; Covino, R.; Faccioli, P. Dominant folding pathways of a WW domain. *Proceedings of the National Academy of Sciences of the United States of America* **2012**, *109* (7), 2330–2335. DOI: 10.1073/pnas.1111796109.

(31) Dave, K.; Jäger, M.; Nguyen, H.; Kelly, J. W.; Gruebele, M. High-Resolution Mapping of the Folding Transition State of a WW Domain. *Journal of molecular biology* **2016**, *428* (8), 1617–1636. DOI: 10.1016/j.jmb.2016.02.008.

(32) Fuller, A. A.; Du, D.; Liu, F.; Davoren, J. E.; Bhabha, G.; Kroon, G.; Case, D. A.; Dyson, H. J.; Powers, E. T.; Wipf, P.; Gruebele, M.; Kelly, J. W. Evaluating beta-turn mimics as beta-sheet folding nucleators. *Proceedings of the National Academy of Sciences of the United States of America* **2009**, *106* (27), 11067–11072. DOI: 10.1073/pnas.0813012106.

(33) Han, W.; Schulten, K. Characterization of folding mechanisms of Trp-cage and WW-domain by network analysis of simulations with a hybrid-resolution model. *The journal of physical chemistry. B* **2013**, *117* (42), 13367–13377. DOI: 10.1021/jp404331d.

(34) Jäger, M.; Nguyen, H.; Crane, J. C.; Kelly, J. W.; Gruebele, M. The folding mechanism of a beta-sheet:The WW domain. *Journal of molecular biology* **2001**, *311* (2), 373–393. DOI: 10.1006/jmbi.2001.4873.

(35) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters* **1999**, *314* (1-2), 141–151. DOI: 10.1016/S0009-2614(99)01123-9.

(36) Otte, L.; Wiedemann, U.; Schlegel, B.; Pires, J. R.; Beyermann, M.; Schmieder, P.; Krause, G.; Volkmer-Engert, R.; Schneider-Mergener, J.; Oschkinat, H. WW domain sequence activity relationships identified using ligand recognition propensities of 42 WW domains. *Protein science : a publication of the Protein Society* **2003**, *12* (3), 491–500. DOI: 10.1110/ps.0233203.

(37) Fowler, D. M.; Araya, C. L.; Fleishman, S. J.; Kellogg, E. H.; Stephany, J. J.; Baker, D.; Fields, S. High-resolution mapping of protein sequence-function relationships. *Nature methods* **2010**, *7* (9), 741–746. DOI: 10.1038/nmeth.1492.

(38) Kato, Y.; Nagata, K.; Takahashi, M.; Lian, L.; Herrero, J. J.; Sudol, M.; Tanokura, M. Common mechanism of ligand recognition by group II/III WW domains: redefining their functional classification. *The Journal of biological chemistry* **2004**, *279* (30), 31833–31841. DOI: 10.1074/jbc.M404719200.

(39) Tzul, F. O.; Vasilchuk, D.; Makhatadze, G. I. Evidence for the principle of minimal frustration in the evolution of protein folding landscapes. *Proceedings of the National Academy of Sciences of the United States of America* **2017**, *114* (9), E1627-E1632. DOI: 10.1073/pnas.1613892114.

(40) Lim, S. A.; Hart, K. M.; Harms, M. J.; Marqusee, S. Evolutionary trend toward kinetic stability in the folding trajectory of RNases H. *Proceedings of the National Academy of Sciences of the United States of America* **2016**, *113* (46), 13045–13050. DOI: 10.1073/pnas.1611781113.

(41) Lim, S. an; Bolin, E. R.; Marqusee, S. Tracing a protein's folding pathway over evolutionary time using ancestral sequence reconstruction and hydrogen exchange. *eLife* **2018**, *7.* DOI: 10.7554/eLife.38369.

(42) Sali, A.; Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *Journal of molecular biology* **1993**, *234* (3), 779–815. DOI: 10.1006/jmbi.1993.1626.

(43) Schymkowitz, J.; Borg, J.; Stricher, F.; Nys, R.; Rousseau, F.; Serrano, L. The FoldX web server:An online force field. *Nucleic acids research* **2005**, *33* (Web Server issue), W382-8. DOI: 10.1093/nar/gki387.

(44) Zhang, C.; Liu, S.; Zhou, Y. Accurate and efficient loop selections by the DFIRE-based all-atom statistical potential. *Protein science : a publication of the Protein Society* **2004**, *13* (2), 391–399. DOI: 10.1110/ps.03411904.

(45) Prilusky, J.; Felder, C. E.; Zeev-Ben-Mordehai, T.; Rydberg, E. H.; Man, O.; Beckmann, J. S.; Silman, I.; Sussman, J. L. FoldIndex:A simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics (Oxford, England)* **2005**, *21* (16), 3435–3438. DOI: 10.1093/bioinformatics/bti537.

(46) Pucheta-Martinez, E.; D'Amelio, N.; Lelli, M.; Martinez-Torrecuadrada, J. L.; Sudol, M.; Saladino, G.; Gervasio, F. L. Changes in the folding landscape of the WW domain provide a molecular mechanism for an inherited genetic syndrome. *Scientific reports* **2016**, *6*, 30293. DOI: 10.1038/srep30293.

TOC Figure