PLOS ONE

RESEARCH ARTICLE

Transferable Coarse-Grained Potential for *De Novo* Protein Folding and Design

Ivan Coluzza*

Faculty of Physics, University of Vienna, Vienna, Austria

*ivan.coluzza@univie.ac.at

Abstract

Protein folding and design are major biophysical problems, the solution of which would lead to important applications especially in medicine. Here we provide evidence of how a novel parametrization of the Caterpillar model may be used for both quantitative protein design and folding. With computer simulations it is shown that, for a large set of real protein structures, the model produces designed sequences with similar physical properties to the corresponding natural occurring sequences. The designed sequences require further experimental testing. For an independent set of proteins, previously used as benchmark, the correct folded structure of both the designed and the natural sequences is also demonstrated. The equilibrium folding properties are characterized by free energy calculations. The resulting free energy profiles not only are consistent among natural and designed proteins, but also show a remarkable precision when the folded structures are compared to the experimentally determined ones. Ultimately, the updated Caterpillar model is unique in the combination of its fundamental three features: its simplicity, its ability to produce natural foldable designed sequences, and its structure prediction precision. It is also remarkable that low frustration sequences can be obtained with such a simple and universal design procedure, and that the folding of natural proteins shows funnelled free energy landscapes without the need of any potentials based on the native structure.

Citation: Coluzza I (2014) Transferable Coarse-Grained Potential for *De Novo* Protein Folding and Design. PLoS ONE 9(12): e112852. doi:10.1371/journal.pone.0112852

Editor: Yang Zhang, University of Michigan, United States of America

Received: July 23, 2014

Accepted: October 20, 2014

Published: December 1, 2014

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. All relevant data are within the paper.

Funding: We acknowledge support from the Austrian Science Fund (FWF) project P23846-N16. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The author has declared that no competing interests exist.

Introduction

Computer simulations of the protein folding process have in the last ten years reached amazing level of description and accuracy [1–16]. The power of the computers and the understanding of the physics that governs folding allows now for a large screening of the experimental data for instance collected in the Protein Data Bank [17]. From a theoretical point of view a successful approach is

the"minimal frustration principle" (MFP) [18–20] in which protein folding is described as a downhill sliding process in a low frustration energy landscape ("funnelled" shaped) towards the native state. While MFP has been proven for lattice heteropolymers [19, 21–27], in more realistic protein representations a residual frustration which prevents the systematic prediction of the native structure of natural sequences is often observed. Off-lattice instead MFP is used as a main justification for the use of structure-based potentials such as the GO [28] and elastic models [29]. In fact, there is still space for development of transferable models that are capable of systematic associating the experimentally determined native structure to natural sequence. Surprisingly, with the exception of few notable examples [30–38], it has also been extremely difficult to artificially construct sequences capable of folding into given target protein structures. The group of David Baker [38] introduced a novel procedure to select sequences with low frustration capable of correctly refolding in vitro to their target structure with a success rate between 8% and up to 40% of the total trials. In their work the authors have introduced a set of rules for the design of the local amino acids interactions to disfavour non-native states. After many iterations, a refolding calculation filters out about 90% of the initial sequences that are found not to have a funnelled energy landscape. The complexity of Baker's procedure demonstrates that is not easy to produce sequences with low frustration.

Here we present a novel protein model where low frustration folding is observed both for natural and designed sequences, the latter obtained without the need of negative design. The novel model is obtained from the optimization of the residue-residue and residue-solvent interaction energy terms under the condition that a large number of sequences designed for 125 test proteins are equal to the corresponding natural sequences. As a result, designed sequences with our model are for several properties comparable to natural ones and fold with a low frustration free energy landscape. We additionally demonstrated that for 15 additional randomly selected proteins, notoriously difficult to fold [39, 40], the natural sequences correctly refolded to their corresponding native structures with a remarkable precision between 2.5 and 5 Å. In other words both quantitative protein design and folding are possible simultaneously. We anticipate that our methodology will have direct application for protein design and structure prediction, but also we expect that it will become a reference point for the development of alternative protein models. For instance, a more or less accurate description can be obtained by adding or removing details from our model, under the condition that the maximum valence principle remains satisfied.

Materials and Methods

Recently we have presented many results that point to the existence of a "maximum valence principle" (MVP) [41–43], according to which for a heteropolymer to be designable and foldable it is sufficient that chain is decorated with directional (low valence) interactions that shape the configurational space. In

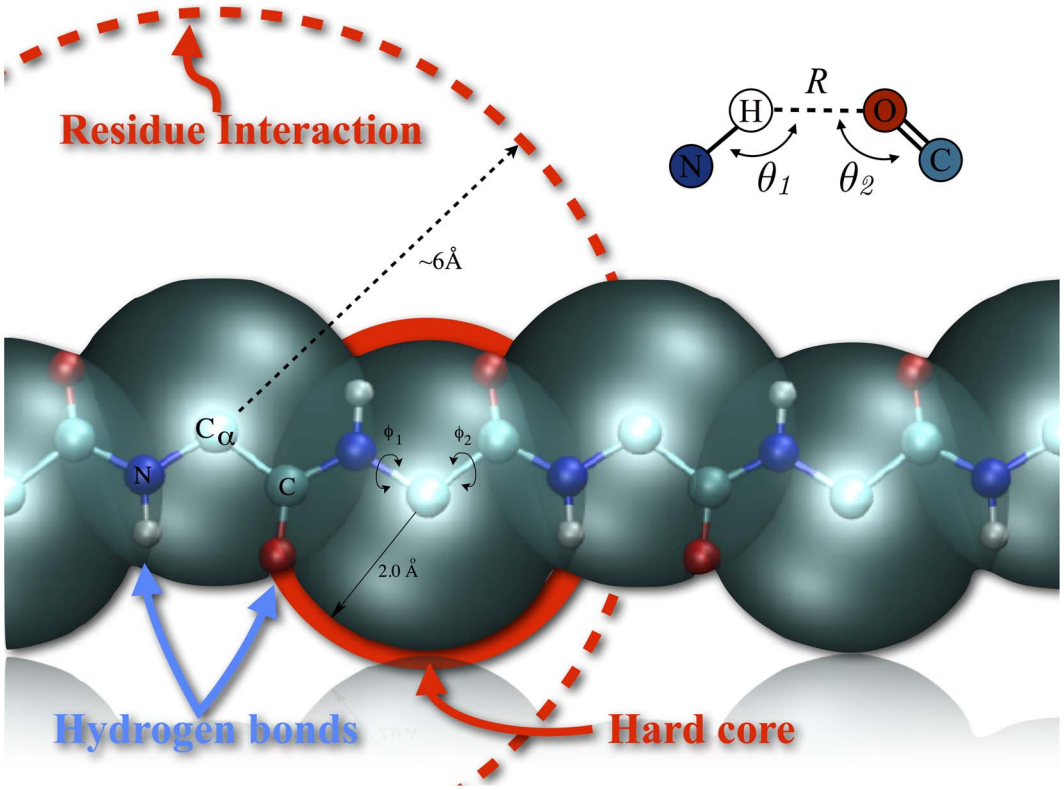

Figure 1. Real-space representation of the backbone of the Caterpillar model. The large blue sphere represent the self-avoidance volume $R_{HC} = 2.0$A of the C_α atoms, while the interaction radius of each residue is represented by the large dashed circle or radius 6 Å (see Eq. S2 in Methods). The H and O atoms interact through a 10–12 Lennard-Jones potential tuned with a quadratic orientation term that selects for alignment of the C, H, O, and N atoms involved in a bond (see top right inset and Eq. S1 in Methods). The backbone fluctuates only around the torsional angles ϕ_1 and ϕ_2.

doi:10.1371/journal.pone.0112852.g001

the case of proteins we have shown (Caterpillar model [43]) that the minimum set of directional interactions translates into the combination of just the backbone molecular geometry and the backbone hydrogen bond interactions (see Fig. 1).

In what follows we will show that by optimizing the interactions under the condition that natural and designed sequences are the same at constant amino acid composition for a large set of proteins, we will also quantitatively predict the folded structures of natural and designed sequences with similar accuracy. This is possible because our model includes the correct set of interactions that satisfy the MVP and, accordingly, the design procedure [43] alone is capable of predicting if a sequence, either natural or artificial, will fold to the target structure. Since we cannot model the particular evolutionary pressure that determined the natural amino acid composition, we chose to keep it constant. Such pressure could be due to many factors such as the particular function of the protein or the difficulty of synthesizing each amino acid type. The ansatz of this work is that folding and design can occur also outside such conditions and that is possible to design a foldable artificial protein from an infinite bath of amino acids. Hence, the above evolutionary pressure is taken into account by fixing the composition to the

natural one. The optimization scheme that we used is the maximum entropy principle (MEP) already tested for proteins by Seno et al. [44]. MEP states that the more information is used to model a system the lower the associated entropy will be [45]. Hence, in order to find the optimal parameters that require the least amount of information, all is needed is to maximize the entropy associated with the probability of observing a given protein $P(S_i,\Gamma_j)$, where S_i indicates to the sequence and Γ_j the three dimensional structure, under the sequence similarity and normalization constraints defined in work of Seno et al. [44]. The derivation follows closely the one used in the work of Seno et al. [44] (the full derivation is in the Supplemental Material together with the details about the model and simulations techniques) and we determined that the entropy maximum corresponds to the values of the model parameters (ϵ, E_{HOH}, and $\mathbf{\Omega}$ in Eq. 1) at which the amino acid hydrophobic/hydrophilic (HP) profile [46] and the interaction energy of each residue with all other are simultaneously equal to the natural ones:

$$\begin{aligned} F_{\mathrm{score}} = & \sum_{j}^{N_{\mathrm{Prot}}} \sum_{k}^{N_j} \left(\sum_{i}^{N_{\mathrm{Seq}}} P(S_i,\Gamma_j) E_{\mathrm{Sol}}^{ik} - E_{\mathrm{Sol}}^{\mathrm{Real}_{jk}} \right)^2 \\ & + \sum_{j}^{N_{\mathrm{Prot}}} \sum_{k}^{N_j} \left(\sum_{i}^{N_{\mathrm{Seq}}} P(S_i,\Gamma_j) \gamma_k^i - \gamma_k^{\mathrm{Real}_j} \right)^2 \\ & + E_{\mathrm{Shannon}} \sum H(\epsilon) \log H(\epsilon) \end{aligned} \tag{1}$$

where the index i runs over the N_{Seq} designed sequences for each protein j of length N_j, the E_{Sol} is the hydrophobicity scale of each residue (see Eq.S3 in the SM), while the γ_k^i's are the contribution to the total energy of each residue calculated within the Caterpillar model. The last term instead guarantees that the Shannon entropy associated to the matrix elements ϵ_{kl} (H are the histograms) is maximized to avoid an uniform matrix. We phenomenologically determined $E_{\mathrm{Shannon}} = 8.0$ for the scaling term to be a good value. Note that here and in the following, energies are given in units of $k_{\mathrm{B}}T_{\mathrm{Ref}}$, where T_{Ref} is a reference temperature that sets the scale of the interactions, hence all simulation temperatures are given in units of T_{Ref}. It is important to stress that T_{Ref} is not necessarily the folding temperature or the environment temperature, but all the energies can be rescaled to have T_{Ref} matching the physical temperature. In fact, in what follows we will show that all proteins studied fold approximately at the same temperature, one could think to rescale the energies to set the folding temperature to the one observed in nature. A schematic representation of the algorithm is reported in Fig. 2.

To the best of our knowledge our work is the first of his kind to optimize the model parameters by reducing the differences between natural and designed sequences and, thanks to the MVP, is the simplest (in terms of the number of

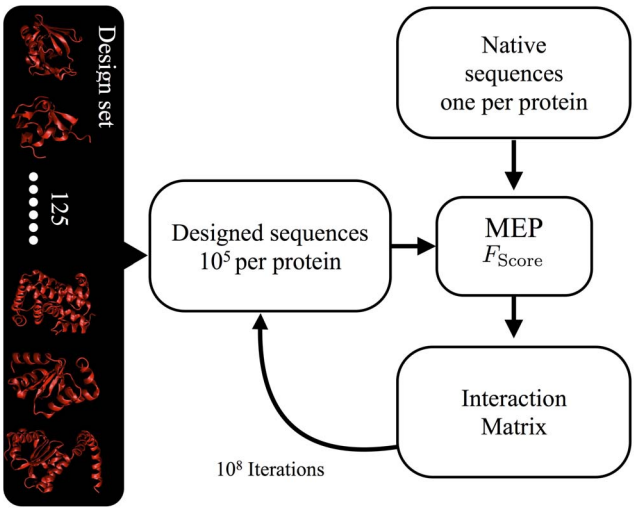

Figure 2. Schematic representation of the MEP algorithm. For a trial set of the ϵ, E_{HOH}, and Ω parameters and for each protein in the training set a large number (10^5) of sequences with composition fixed to the natural one are generated following the design scheme in the SM. The scoring function F_{score} (Eq.S16 in the SM) is then evaluated and the trial parameters are accepted or reject according to a Metropolis like scheme. New parameter sets are generated at each iteration, and the sequences of the proteins in the training set are re-designed by 10^5 simple pair residue swapping moves, which are accepted or rejected according to a standard Metropolis algorithm with the energy defined in Eq.S4 (see SM). During each design iteration, the HP and energy profiles (Eq.S16 in the SM) are averaged over the observed sequences weighted by their Boltzmann weight. The averaging guarantees that the profiles are calculated over the most probable sequences that, as we showed previously [43], are robust against mutations and are more thermally stable. After $\sim 10^8$ iterations the interaction parameters converged to their final values: $\Omega = 21.0 \pm 0.5$ and $E_{\mathrm{HOH}} = 0.015 \pm 0.001$, and the residue-residue ϵ interaction parameters which are listed in Tab. S1 of the SM.

doi:10.1371/journal.pone.0112852.g002

parameters needed) to successfully and quantitatively reproduce both sequences and structures of natural proteins to high precision.

Results

Parameters Optimization

We began by selecting a protein training set from the Protein Data Bank (PDB) [17], which includes all the proteins that obey the following conditions: a X-Ray structural resolution below 1.5 Å, are made of single chains of length ranging from 20 to 200 residues, and do not contain any DNA or RNA. According to the stated conditions we selected 125 proteins (see Tab. S2 of the SM for the complete list of the PDB id's). It is important to stress that we did not select for specific experimental conditions, in particular pH and temperature during the measurements fluctuate significantly among the proteins in the set. In Fig. 3 and Fig. S3 we plot the comparison of the natural to the designed sequences, the latter obtained with the MEP optimized interaction parameters. The plot shows strong correlation (>0.9) between the total energy of the designed (abscissa) and natural (ordinate) sequences, and between the profiles of the residue the HP profiles and the energy contribution (Fig. 3 top and bottom insets and Fig. S3). *Overall we can*

conclude that, for all 125 proteins in the training set, the designed proteins and natural proteins are equally compatible sequences to their respective target structures, strongly suggesting that our procedure may now be used to design realistic protein sequences. We applied the MEP derived parameters to design 15 randomly selected from independent training sets [39, 40], and characterized by different secondary and tertiary motives. The top five resulting sequences for each target structures are listed in Tab. S4 of the SM. It is important to stress for this design we relaxed the constraint on the amino acid composition used during the optimization. Hence, the folding of the designed sequences does not depend on the previous knowledge of the natural amino acid composition, nevertheless the amino acids composition of the artificial sequences is similar to the natural one (see Tab. S3). It has to be said that the artificial sequences appear unusual with repeats of the same amino acid (e.g. for 1gab *WDDMIIRRRRFVVYYLWGSMTAEVEAEKGTNGFYYHHHD-FGTKKKAQQQSNNL*). Such repeats could be due to the approximations of the model, however it is important to remember that we did not include in the design any information about the function of the protein. In fact there is no reason to expect that natural sequences are the only one capable of folding, and we want to stress again that additional constraints applied during the design procedure would dramatically reduce the volume of the sequence space [47] reducing the probability of repeats. We believe the latter to be the main cause of the repeats and we tested this hypothesis forcing into the design procedure and additional constraint expressly rejecting mutations that would result into a repetition of the same amino acid for 3 residues forward and backwards. The resulting sequences are listed in Tab. S6 of the SM. We were surprised to find that sequences with energy comparable to the unconstrained ones had a lower number of permutations ($log(N_P) \sim 107$ instead of 108). Nevertheless, the now more reasonable looking sequences folded in a similar fashion with respect to their unconstrained alternatives (see Fig. $\tilde{S}4$). Finally, we will show below the model is capable of refolding also several natural sequences, demonstrating that in the model the presence of repeats is not necessary to stabilize natural protein structures. It is important to stress that during the last step of the MEP optimization the fewer sequences generated with fixed composition do not present the repeating patterns (see Table S5). However, this is an interesting problem and deserves a dedicated study that is beyond the objective of this work. Objective of ongoing research is also to experimentally test whether such sequences are capable of folding to the predicted target structures.

Protein design

In order to apply the model to the folding of both designed and natural sequences, we need to balance the residue energy term with the backbone hydrogen bond term (parameter α in Eq.S4 in the SM). The energies can be rescaled by choosing the value of α for which designed sequences fold best to their target structures [43]. Hence, we selected four designed sequences from Tab. S4 (PDB ids 2l09, 3mx7,chain A of 3obh, and 1qyp), and for each sequence we performed a refolding simulation (see SM) with different values of the rescaling parameter α in

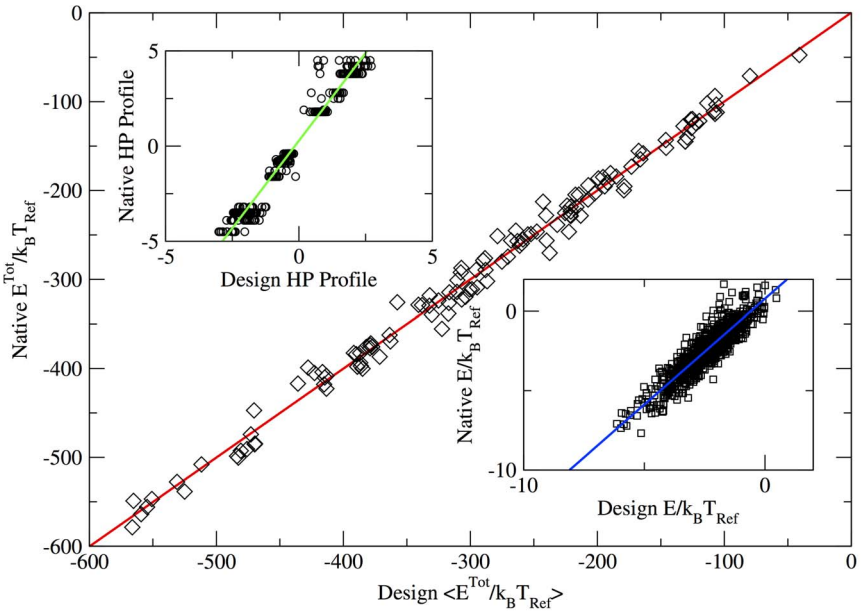

Figure 3. Comparison between the total residue energy $<E^{Tot}/k_BT_{Ref}>$ (Eq.S4 in SM) averaged over all the 10^5 designed sequences per target(abscissa) and the same energy calculated over the native sequence of same target (ordinate). Each point corresponds to one protein in the data set and shows a strong linear trend verified by the fit (red line) with a correlation coefficient of ~0.995 and a slope of ~1.000 indicating that two energies are perfectly correlated. In the insets we show the comparison of the HP profiles (top left) and interaction energy E/k_BT_{Ref} of each residue with all other (bottom right), this time each point corresponds to a single residue of each test protein. In both cases the data follow a remarkable linear trend (fits in green and blue lines respectively), and a positive correlation close to unity. For the HP profiles the correlation coefficient (~0.98) indicates that when in natural proteins we find an hydrophobic residue also the design procedure will put one and vice versa. While the correlation coefficient (~0.90) of E/k_BT_{Ref} demonstrates that each natural residue has a very similar contribution to the total energy compared to the designed ones. A perfect match cannot be expected since natural sequences might have experience a selection pressure influenced by interactions not represented in the model, different environmental conditions or simply unknown functional requirements. Nevertheless the accordance is remarkable.

doi:10.1371/journal.pone.0112852.g003

the range [0.05 to 1.0]. The best value of $\alpha = 0.10 \pm 0.01 k_BT_{Ref}$ was the one for which all four proteins folded closer and smoother to the native state. In Fig. 4 we plot the refolding free energy $F(\text{DRMSD})/k_BT_{Ref}$ as a function of the distance root mean square displacement (DRMSD, see Appendix DMRSD and Fig. S2 of the SM), obtained with the best energy value for $\alpha = 0.10 \pm 0.01 k_BT_{Ref}$ for the four target proteins below the folding temperatures (estimated to be $T_F \approx 0.22$ for all proteins see Fig. S1 in the SM for details). The plot shows for each protein a funnelled profile with a global minimum very close to the respective target structure (DRMSD $\in [1.5-2.0]$ Å). So at least below the folding temperature the proteins seems to follow a downhill process. This observation would need a verification with a study of the folding dynamics. *The refolding free energy profiles shown in Fig. 4 prove that realistic protein sequences with low frustration folding free energy landscapes can now be designed with a straightforward positive design scheme.*

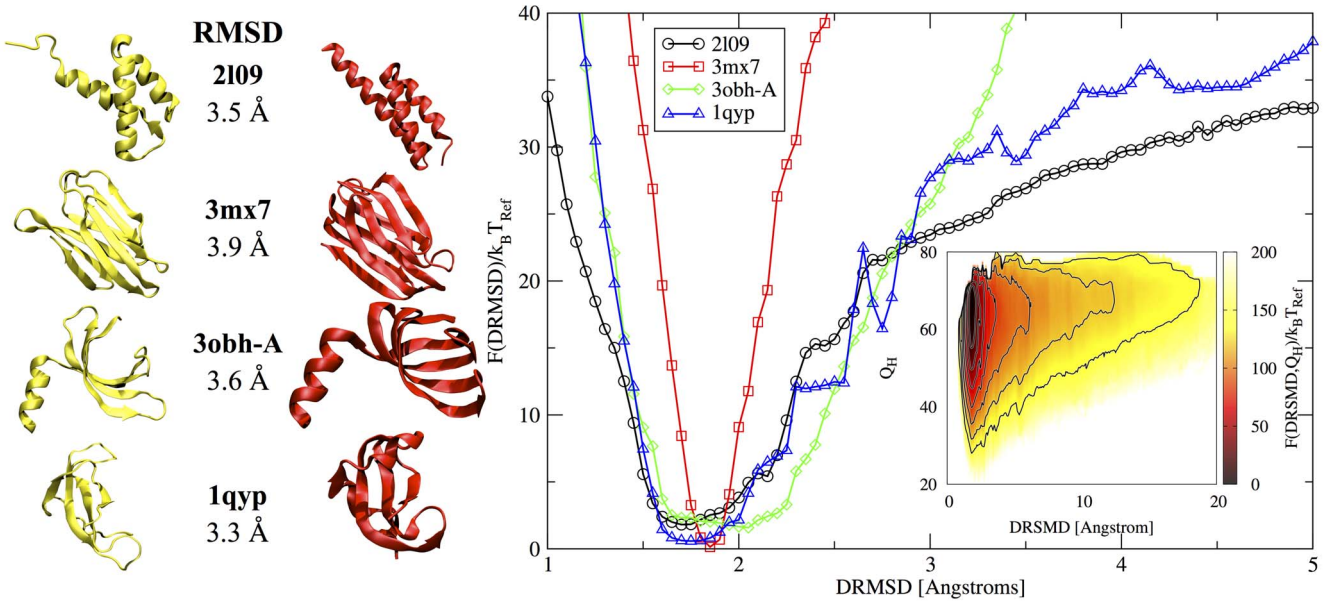

Figure 4. Folding free energy landscape $F(\mathrm{DRMSD})/k_BT_{\mathrm{Ref}}$ as a function of DRMSD of the four designed proteins (PDB ids 2l09, 3mx7, chain A of 3obh, and 1qyp). All profiles have a global minimum around 1.5 and 2 Å DRMSD with a smooth funnelled shape. Due the approximations present in the model and to thermal fluctuations is shifted with respect to $\mathrm{DRMSD}=0$ (note that to the value $\mathrm{DRMSD}=0$ of each profile will correspond a different native structure). Because of the definition of DRMSD, the smaller the value the fewer are the possible structures that can have this value of DRMSD. Ultimately, $\mathrm{DRMSD}=0$ is possible only for the target structure itself. The funnelled profiles with single minimum implies that both an ensemble of arrested structures and a single alternative fold are less stable compared to the desired configuration. In the bottom right inset we plot the folding free energy landscape $F(\mathrm{DRMSD},Q_H)/k_{\mathrm{B}}T_{\mathrm{Ref}}$ for 3mx7 as a function of both the DRMSD and the number of hydrogen bonds Q_H, to give a visual example of the funnel nature of the folding landscapes. On the left we compare the experimentally determined structures (in yellow) with a typical folded conformation selected as the sampled configurations with the lowest energy at the free energy minimum (in red). The RMSD value is indicated in the middle. The structures were aligned using the *RMSD calculator* tool in VMD [52], while the secondary structure elments where identified with STRIDE [53].

doi:10.1371/journal.pone.0112852.g004

We now have obtained the optimized parameters for our model: $\alpha=0.10\pm0.01k_BT_{\mathrm{Ref}}$, $\Omega=21.0\pm0.5$, $E_{\mathrm{HOH}}=0.015\pm0.001$ and for the ϵ see Tab. S1 in the SM.

Refolding of natural sequences

The next logical step is to asses the behaviour of the model when refolding natural sequences and prove that folding as well can be performed to a quantitative level with the model. For this we randomly selected 15 proteins known to be difficult to fold (from Tsai et al. [39] and from the 9th edition of the well known Critical Assessment of Techniques for Protein Structure Prediction [40]) and we performed folding simulations of their natural sequences. The results are plotted in Fig. 5(a) and 5(b), where we have superimposed all the computed free energy profiles. Although, the details of each profile might not be clearly visible, a first fundamental feature is apparent, namely the concentration of the free energy minima in the region between 1.5 and 2 Å DRMSD which remarkably is also the same regions observed for the design proteins. A second important result is the funnel shape common to all free energy profiles providing definite proof of the capability of the model of capturing the low frustration folding of natural proteins

with a rather high precision. In fact when the predicted conformations of the folded states are compared to the experimentally determined structures, the two overlapped with a precision between 2.4 and 4.1 Å RMSD (see top inset of Fig. 5(a)) which is surprisingly accurate especially considering the simplicity of the model. An alternative comparison of the refolded structures to their native targets is reported in Tab. S7 produced with the ''MaxCluster'' program from Alex Herbert (http://www.sbg.bio.ic.ac.uk/maxcluster/index.html) and the TM-scoring function introduced by Zhang et al. [48, 49]. It is important to note that the configurations with the lowest energy are not necessarily equal to the ones corresponding to the minimum of the free energy, however, in most cases, they are very similar. This is due to the strong directional nature of the hydrogen bonds which makes them very sensitive to thermal fluctuations. As a consequence, there are isolated structures that might have a lower energy but are not very stable at finite temperature. We would like to stress that, since the native sequences fold in a similar fashion compared to the designed ones and we did not observe the native sequences themselves as an outcome of the design process (see Tab. S4–S6), we could speculate that, at least within the Caterpillar model, the space of folding sequences is much wider than the one comprising only of natural sequences.

Discussion

To the best of our knowledge our coarse-grained protein model is the simplest, in terms of the number of parameters needed, with a transferable energy function capable of achieving such precision for the prediction of the native folded structures. Also it is one of the very few models that allows for both quantitative proteins design and folding, the latter demonstrated by free energy calculations. It is remarkable that low frustration sequences can be obtained with such a simple and universal design procedure, and that the folding of natural proteins shows funnelled free energy landscapes without the need of any potentials based on the native structure [50].

Although, the artificial sequences present some unnatural features like repetitions of some amino acids, the sequences designed with a natural amino acid composition share many features with the natural occurring ones, and the native structures of the latter are correctly predicted by our model. Hence, we expect that our designed proteins (see Tab. S4), once synthesized, may fold to the structures used as design targets, which may also represent the ultimate and most important test of our methodology. We hope that our methodology will become an useful tool in experiments requiring alterations of natural proteins, or the total redesign of target protein structures. Of course, constraints on the composition can always be applied to the design procedure with no major changes in the procedure. Moreover, the prediction power of the model gives us high confidence that our design methodology may be directly used to tackle important open problems of drug design, or used in a multi-scale approach where the results from our model

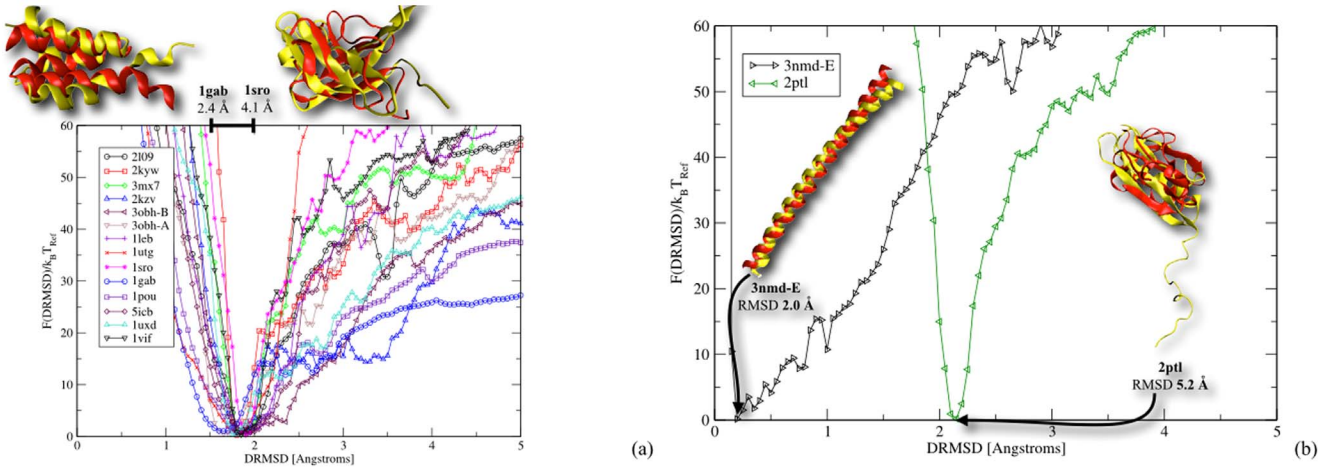

Figure 5. Folding free energy landscape $F(\mathrm{DRMSD})/k_BT_{\mathrm{Ref}}$ of the 15 proteins set selected to test the accuracy of the MEP optimized parameters. The profiles have a common funnel shape and show a clustering of the free energy minima in the region 1.5 and 2 Å DRMSD consistent with the results obtained for designed sequences. In b) we plot the free energies for proteins with the worst (2ptl) and the best (3nmd-E) distance of the folded structure from the native one. For the latter the free energy profile shows a minimum remarkably close to the native state probably due to the highly simplified structure of protein 3nmd-E. The minimum of 2ptl, on the other hand, is located further away from the low DRMSD values than the other proteins. This apparent discrepancy is due to the definition of the DRMSD which includes the contribution from the C_α atoms located in the long unstructured tail from the residue 1 to 18. Since the probability of observing that particular conformation in solution is very low, it follows that the particular realization of the native structure has a large entropy penalty. However if we measure the overlap ignoring the contribution from the tail we see that the predicted structure of the protein core is again reasonably close to the experimentally determined one (≈5.2 Å RMSD). In the insets we compare the experimental structures (in yellow) superimposed to the equilibrium configurations (in red), and we show that the proteins refolded with a precision between 2.4 and 4.1 Å RMSD.

doi:10.1371/journal.pone.0112852.g005

could be refined with a more accurate but also a computationally more expensive protein model.

Finally, this work not only extends our previous results obtained with the Caterpillar model, but also strengthens the connection among all our work on lattice heteropolymers and protein unrelated systems such as patchy polymers [41, 51]. The success of the same design strategy for all these systems demonstrates that the maximum valence principle is a sufficient condition to satisfy for the generalized design of low frustration sequences and the prediction of their proper native state.

Acknowledgments

We would like to thank Peter van Oostrum, Barbara Capone, Francisco Martinez-Veracoechea, Angelo Cacciuto and Christoph Dellago for fruitful discussions and a critical reading of the manuscript. All simulations presented in this paper were carried out on the Vienna Scientific Cluster (VSC). The images depicting protein structures were made with VMD/NAMD/BioCoRE/JMV/other software support. VMD/NAMD/BioCoRE/JMV/is developed with NIH support by the Theoretical and Computational Biophysics group at the Beckman Institute, University of Illinois at Urbana-Champaign.

Supporting Information

Figure S1. Folding free energy landscape $F(\mathrm{DRMSD})/k_{\mathrm{B}}T_{\mathrm{Ref}}$ as a function of DRMSD of the designed protein PDB ids 1CTF close to the folding temperature.
doi:10.1371/journal.pone.0112852.S001 (TIFF)

Figure S2. On the left: correlation plot between the DRMSD and the RMSD collective variable. The estimated correlation coefficient from a linear regression fitting (in red) is $\approx$0.8 which increases to $\approx$0.98, if we exclude the configurations for values of DRMSD $<$1.5 $\AA$ which is below the model resolution, indicating that the free energy profile should be qualitatively similar if the states are projected over RMSD instead of DRMSD. On the right: Free Energy folding profile of the protein 3NMD-E projected over the collective variables DRMSD and RMSD. The profiles are not identical because the RMSD is more sensitive to local distortions of the protein with respect to the DRMSD. This is also demonstrated by the wider free energy minimum which reflects the thermal fluctuations. However, overall the qualitative shape of the profiles is very similar with between each other in particular since both have a clear global free energy minimum.
doi:10.1371/journal.pone.0112852.S002 (TIFF)

Figure S3. Correaltion between designed and real E_{sol} profiles. The correlation coefficient has been estimated from a linear regression fitting (in red) to be be very high $\approx$0.98.
doi:10.1371/journal.pone.0112852.S003 (TIFF)

Figure S4. Folding free energy landscape $F(\mathrm{DRMSD})/k_{\mathrm{B}}T_{\mathrm{Ref}}$ as a function of DRMSD of the four designed proteins (PDB ids 2l09, 3mx7, chain A of 3obh, and 1qyp). All profiles have a global minimum around 1.5 and 2 Å DRMSD with a smooth funnelled shape. Due the approximations present in the model and to thermal fluctuations is shifted with respect to DRMSD = 0 (note that to the value DRMSD = 0 of each profile will correspond a different native structure). Because of the definition of DRMSD, the smaller the value the fewer are the possible structures that can have this value of DRMSD. Ultimately, DRMSD = 0 is possible only for the target structure itself. The funnelled profiles with single minimum implies that both an ensemble of arrested structures and a single alternative fold are less stable compared to the desired configuration. In the bottom right inset we plot the folding free energy landscape $F(\mathrm{DRMSD},Q_H)/k_{\mathrm{B}}T_{\mathrm{Ref}}$ for 3mx7 as a function of both the DRMSD and the number of hydrogen bonds Q_H, to give a visual example of the funnel nature of the folding landscapes. On the left we compare the experimentally determined structures (in yellow) with a typical folded conformation selected as the sampled configurations with the lowest energy at the free energy minimum (in red). The RMSD value is indicated in the middle.
doi:10.1371/journal.pone.0112852.S004 (TIFF)

Table S1. Optimized values of the residue-solvent ϵ_{Sol} and residue-residue $\epsilon(S_k)(S_l)$ interaction parameters. The uncertainty on the values is $\approx \pm 0.01$.
doi:10.1371/journal.pone.0112852.S005 (PDF)

Table S2. List of PDB id's used as training set for the maximum entropy parameters optimization.
doi:10.1371/journal.pone.0112852.S006 (PDF)

Table S3. Comparison of the average composition of the designed sequences and the natural sequences used in the parameter optimization. It is important that since we do not model Cys-Cys bond and the Proline rigid bond we have excluded them from the design alphabet. This is why the frequency associated to those amino acids is zero in the designed sequences. We are currently working on implementing such special cases in the Caterpillar model. We have highlighted in bold the amino acids types with the largest discrepancies namely: Histidine, Methionine, Tryptophan, Tyrosine. Such amino acids are know to be the one with the lowest appearance frequency in nature. Since we did not impose any restriction on the design procedure over the relative abundance of amino acids in nature it is not surprising to find the largest discrepancies in the composition for such amino acids.
doi:10.1371/journal.pone.0112852.S007 (PDF)

Table S4. Designed sequences.
doi:10.1371/journal.pone.0112852.S008 (PDF)

Table S5. Sequences obtained during the last step of the matrix optimization procedure. The amino acid composition is identical for all sequences and the first is the natural sequences taken from the pdb file.
doi:10.1371/journal.pone.0112852.S009 (PDF)

Table S6. Designed sequences under the additional constraint that local repetition of up to 5 residues are forbidden.
doi:10.1371/journal.pone.0112852.S010 (PDF)

Table S7. Summary of the refolded structures with the natural sequences. The DRMSD value is taken form the minimum of the folding free energy (see Fig. 5), while the Overlap, the gRMSD and the TM-score are calculated using the Max Cluster program from Alex Herbert (http://www.sbg.bio.ic.ac.uk/maxcluster/index.html). The Overlap is a measure of percentage of matched structural elements between the native and the refolded structures. The gRMSD and TM-score are calculated over the overlapping structural elements. TM-score was defined by Zhang et al. [48, 49]. The low overlapping value for 2ptl and 1vif are due to the unstructured sections of the proteins.
doi:10.1371/journal.pone.0112852.S011 (PDF)

Text S1. Supplemental Material containing the details about the model and simulations techniques together with the derivation of the scoring function with the Maximum Entropy Principle.
doi:10.1371/journal.pone.0112852.S012 (PDF)

Author Contributions

Conceived and designed the experiments: IC. Performed the experiments: IC. Analyzed the data: IC. Contributed reagents/materials/analysis tools: IC. Wrote the paper: IC.

References

1. **Voegler Smith a, Hall CK** (2001) Alpha-Helix Formation: Discontinuous Molecular Dynamics on an Intermediate-Resolution Protein Model. Proteins 44: 344–60.
2. **Pande VS, Baker I, Chapman J, Elmer SP, Khaliq S, et al.** (2003) Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing. Biopolymers 68: 91–109.
3. **Hamelberg D, Mongan J, McCammon JA** (2004) Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. The Journal of chemical physics 120: 11919–11929.
4. **Seibert MM, Patriksson A, Hess B, van der Spoel D** (2005) Reproducible polypeptide folding and structure prediction using molecular dynamics simulations. Journal of molecular biology 354: 173–83.
5. **Tozzini V** (2005) Coarse-grained models for proteins. Current Opinion in Structural Biology 15: 144–50.
6. **Henzler-Wildman K, Kern D** (2007) Dynamic personalities of proteins. Nature 450: 964–72.
7. **Mu Y, Gao YQ** (2007) Effects of hydrophobic and dipole-dipole interactions on the conformational transitions of a model polypeptide. The Journal of chemical physics 127: 105102.
8. **Huang X, Bowman GR, Pande VS** (2008) Convergence of folding free energy landscapes via application of enhanced sampling methods in a distributed computing environment. The Journal of chemical physics 128: 205106.
9. **Laio A, Gervasio FL** (2008) Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. Reports on Progress in Physics 71: 126601.
10. **Bereau T, Deserno M** (2009) Generic coarse-grained model for protein folding and aggregation. The Journal of chemical physics 130: 235106.
11. **Shaw DE, Maragakis P, Lindorff-Larsen K, Piana S, Dror RO, et al.** (2010) Atomic-level characterization of the structural dynamics of proteins. Science (New York, NY) 330: 341–6.
12. **Zhang C, Ma J** (2010) Enhanced sampling and applications in protein folding in explicit solvent. The Journal of chemical physics 132: 244101.
13. **Ikebe J, Standley DM, Nakamura H, Higo J** (2011) Ab initio simulation of a 57-residue protein in explicit solvent reproduces the native conformation in the lowest free-energy cluster. Protein science: a publication of the Protein Society 20: 187–96.
14. **Lindorff-Larsen K, Piana S, Dror RO, Shaw DE** (2011) How Fast-Folding Proteins Fold. Science 334: 517–520.
15. **Zhang C, Ma J** (2012) Folding helical proteins in explicit solvent using dihedral-biased tempering. Proceedings of the National Academy of Sciences of the United States of America 109: 8139–44.
16. **Kapoor A, Travesset A** (2013) Folding and stability of helical bundle proteins from coarse-grained models. Proteins 81: 1200–11.
17. **Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al.** (2000) The Protein Data Bank. Nucleic acids research 28: 235–42.
18. **Bryngelson JD, Wolynes PG** (1987) Spin glasses and the statistical mechanics of protein folding. Proceedings of the National Academy of Sciences of the United States of America 84: 7524–8.
19. **Shakhnovich EI, Gutin AM** (1993) Engineering of stable and fast-folding sequences of model proteins. Proceedings of the National Academy of Sciences of the United States of America 90: 7195–9.
20. **Wolynes PG, Eaton Wa, Fersht aR** (2012) From the Cover: Chemical physics of protein folding. Proceedings of the National Academy of Sciences 109: 17770–17771.

21. **Gutin AM, Shakhnovich E** (1993) Ground-state of random copolymers and the discrete Random Energy-model. The Journal of Chemical Physics 98: 8174–8177.

22. **Shakhnovich E** (1994) Proteins with selected sequences fold into unique native conformation. Physical Review Letters 72: 3907–3910.

23. **Shakhnovich EI, Gutin AM** (1990) Implications of thermodynamics of protein folding for evolution of primary sequences. Nature 346: 773–5.

24. **Coluzza I, Muller HG, Frenkel D** (2003) Designing refoldable model molecules. Physical Review E 68: 46703.

25. **Coluzza I, Frenkel D** (2004) Designing specificity of protein-substrate interactions. Physical Review E 70: 51917.

26. **Coluzza I, Frenkel D** (2007) Monte Carlo study of substrate-induced folding and refolding of lattice proteins. Biophysical Journal 92: 1150–6.

27. **Abeln S, Frenkel D** (2008) Disordered flanks prevent peptide aggregation. PLoS computational biology 4: e1000241.

28. **Go N** (1983) Theoretical studies of protein folding. Annual review of biophysics and bioengineering 12: 183–210.

29. **Atilgan aR, Durell SR, Jernigan RL, Demirel MC, Keskin O, et al.** (2001) Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophysical journal 80: 505–15.

30. **Dahiyat BI, Mayo SL** (1997) Probing the role of packing specificity in protein design. Proceedings of the National Academy of Sciences of the United States of America 94: 10172–7.

31. **Desjarlais JR, Handel TM** (1995) De novo design of the hydrophobic cores of proteins. Protein science: a publication of the Protein Society 4: 2006–18.

32. **Desjarlais JR, Handel TM** (1999) Side-chain and backbone flexibility in protein core design. Journal of molecular biology 290: 305–18.

33. **Hellinga HW, Richards FM** (1991) Construction of new ligand binding sites in proteins of known structure. Journal of Molecular Biology 222: 763–785.

34. **Distasio RA, von Lilienfeld OA, Tkatchenko A** (2012) Collective many-body van der Waals interactions in molecular systems. Proceedings of the National Academy of Sciences of the United States of America 109: 14791–5.

35. **Rothlisberger D, Khersonsky O, Wollacott AM, Jiang L, Dechancie J, et al.** (2008) Kemp elimination catalysts by computational enzyme design. Nature 453: 190–U4.

36. **Dahiyat BI** (1997) De Novo Protein Design: Fully Automated Sequence Selection. Science 278: 82–87.

37. **Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, et al.** (2003) Design of a novel globular protein fold with atomic-level accuracy. Science (New York, NY) 302: 1364–8.

38. **Kellogg EH, Lange OF, Baker D** (2012) Evaluation and optimization of discrete state models of protein folding. The journal of physical chemistry B 116: 11405–13.

39. **Tsai J, Bonneau R, Morozov AV, Kuhlman B, Rohl CA, et al.** (2003) An improved protein decoy set for testing energy functions for protein structure prediction. Proteins 53: 76–87.

40. **Kinch LN, Shi S, Cheng H, Cong Q, Pei J, et al.** (2011) CASP9 target classification. Proteins 79 Suppl 1: 21–36.

41. **Coluzza I, van Oostrum PDJ, Capone B, Reimhult E, Dellago C** (2012) Design and folding of colloidal patchy polymers. Soft Matter DOI: 10.1039/c2sm26967h.

42. **Coluzza I, Dellago C** (2012) The configurational space of colloidal patchy polymers with heterogeneous sequences. Journal of Physics: Condensed Matter 24: 284111.

43. **Coluzza I** (2011) A coarse-grained approach to protein design: learning from design to understand folding. PloS one 6: e20853.

44. **Seno F, Trovato A, Banavar J, Maritan A** (2008) Maximum Entropy Approach for Deducing Amino Acid Interactions in Proteins. Physical Review Letters 100: 1–4.

45. **Shannon CE** (1948) A Mathematical Theory of Communication. Bell System Technical Journal 27: 379–423.

46. **Dolittle RF** (1989) In Predictions of Protein Structure and the Principles of Protein Conformation. Springer, 599–623 pp.

47. **Coluzza I, MacDonald JT, Sadowski MI, Taylor WR, Goldstein Ra** (2012) Analytic markovian rates for generalized protein structure evolution. PloS one 7: e34228.

48. **Zhang Y, Skolnick J** (2004) Scoring function for automated assessment of protein structure template quality. Proteins 57: 702–10.

49. **Xu J, Zhang Y** (2010) How significant is a protein structure similarity with TM-score $=0.5$? Bioinformatics (Oxford, England) 26: 889–95.

50. **Go N, Taketomi H** (1978) Respective roles of short-range and long-range interactions in protein folding. Proceedings Of The National Academy Of Sciences Of The United States Of America 75: 559–563.

51. **Coluzza I, van Oostrum PDJ, Capone B, Reimhult E, Dellago C** (2013) Sequence Controlled Self-Knotting Colloidal Patchy Polymers. Physical Review Letters 110: 075501.

52. **Humphrey W, Dalke A, Schulten K** (1996) {VMD} – {V}isual {M}olecular {D}ynamics. Journal of Molecular Graphics 14: 33–38.

53. **Frishman D, Argos P** (1995) Knowledge-based protein secondary structure assignment. Proteins 23: 566–79.