PLOS ONE

RESEARCH ARTICLE

De novo protein structure prediction using ultra-fast molecular dynamics simulation

Ngaam J. Cheung[1,2], **Wookyung Yu**[1,3]*

1 Department of Brain and Cognitive Science, DGIST, Daegu, South Korea, **2** Cavendish Laboratory, Department of Physics, University of Cambridge, Cambridge, United Kingdom, **3** Core Protein Resources Center, DGIST, Daegu, South Korea

* wkyu@dgist.ac.kr

OPEN ACCESS

Citation: Cheung NJ, Yu W (2018) *De novo* protein structure prediction using ultra-fast molecular dynamics simulation. PLoS ONE 13(11): e0205819. https://doi.org/10.1371/journal.pone.0205819

Editor: Yang Zhang, University of Michigan, UNITED STATES

Received: June 2, 2018

Accepted: October 2, 2018

Published: November 20, 2018

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by DGIST start-up fund No. 2018010089 and the Korean Government Ministry of Trade, Industry and Energy N0001822. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

Abstract

Modern genomics sequencing techniques have provided a massive amount of protein sequences, but experimental endeavor in determining protein structures is largely lagging far behind the vast and unexplored sequences. Apparently, computational biology is playing a more important role in protein structure prediction than ever. Here, we present a system of *de novo* predictor, termed *NiDelta*, building on a deep convolutional neural network and statistical potential enabling molecular dynamics simulation for modeling protein tertiary structure. Combining with evolutionary-based residue-contacts, the presented predictor can predict the tertiary structures of a number of target proteins with remarkable accuracy. The proposed approach is demonstrated by calculations on a set of eighteen large proteins from different fold classes. The results show that the ultra-fast molecular dynamics simulation could dramatically reduce the gap between the sequence and its structure at atom level, and it could also present high efficiency in protein structure determination if sparse experimental data is available.

Introduction

In modern biology and medicine, it is a major challenge to determine a protein tertiary structure from its primary amino acid sequence, and it has significant and profound consequences, such as understanding protein function, engineering new proteins, designing drugs or for environmental engineering [1–3]. Nowadays, more and more protein sequences are being produced by genomics sequencing techniques. Despite tremendous efforts of community-wide in structural genomics, protein structures determined by experiments, such as X-ray crystallography, NMR spectroscopy or Cryo-EM, cannot keep the pace with the explosive growth of protein sequences [4]. Since it requires numerous time and relatively expensive efforts, experimental determination of protein structures is lagging behind, and the gap between sequences and structures is widening rather than diminishing [5].

Amino acid sequences contain enough information for specifying their three-dimensional structures [6], thus which provides the principle for predicting three-dimensional structure from its sequence. Accordingly, in the past decades, computational prediction of protein

structures has been a long-standing challenge, and a number of computational methods have been contributed to bridge the gap, which may be able to be reduced or filled if the approaches can provide predictions of sufficient accuracy [5]. As efficient models, template or homology modeling methods [7–9] utilize the similarity of the query sequence (target) to at least one protein of known tertiary structure, and protocols in these methods enable to accurately predict protein three-dimensional conformation from its amino acid sequence. However, template or homology models cannot work if there is no determined structure in the same protein family as that of the query sequence. Only relying on the amino acid sequence and no structural template, *de novo* approaches depend on an effective conformation-searching algorithm and good energy functions to build protein tertiary structures.

Nowadays, *de novo* predictors remain restricted to small proteins, and most of them are extremely difficult to achieve on large proteins because of the vast conformational space and computational bottlenecks [10, 11]. Some of these *de novo* approaches rely on assembling proteins from short peptide fragments, which are derived from known proteins based on the sequence similarity [8, 9]. For example, Rosetta utilizes sequence-similar fragments by searching against three-dimensional structure databases followed by fragment assembly using empirical intermolecular force fields [12]. Although many striking *de novo* advances have been achieved, such methods have worked on smaller proteins that have less than 100 amino acids [13, 14], unfortunately, the *de novo* structure prediction problem is still unsolved and presents a fundamental computational challenge, even for fragment-based methods [13].

Here we describe an approach, termed *NiDelta*, to predict protein tertiary structure from amino acid sequence. *NiDelta* models a protein structure from its amino acid sequence primarily involving three steps: (a) predicting torsional angles (ϕ, ψ) based on the convolutional neural network (CNN); (b) capturing residue contacts based on evolutionary information; and (c) sampling conformation space by ultra-fast Molecular Dynamics simulation.

Materials and methods

In this section, the developed *NiDelta* is described in details. The framework of *NiDelta* is illustrated in Fig 1. As shown, for a given target sequence, *NiDelta* will prepare two main restraints, which are predicted torsion angle and residue-contacts for launching a coarse-grained molecular dynamics (CGMD)—*Upside* [15] for sampling conformation space. As illustrated in the Fig 1, there are two stages to process data: 1) training the *Phsior*, and 2) estimations of residue-contacts. In the stage of building the *Phsior*, we construct a non-redundant sequence data set from RCSB PDB library and culled it through PICSCES [16]. Then, a deep convolutional neural network [17] (termed *Phsior*, a module in Sibe web-server [18] will be trained using the fine-tuned data set (not include the 18 proteins as shown in Table 1). Thereafter, the trained *Phsior* is used to predict torsional angles (ϕ, ψ) of a given query amino acid sequence.

The data set was not used to prepare the MSA. On the other hand, the MSA that is used to infer the residue-contacts was obtained by searching against the UNIREF100 database by HMMER suite (Jackhmmer). Then the obtained MSA will be trimmed and filtered to remove invalid sequences and keep the efficient sequences that enhance the quality of DCA estimation. On the other hand, for the same query sequence, we search it against UNIREF100 database [19] by HMMER [20] to obtain an alignment of multiple sequences. Then the obtained MSA will be trimmed and filtered to remove invalid sequences and keep the efficient sequences that enhance the quality of estimating residue-contacts. Accordingly, residue contacts are inferred from the multiple sequence alignment, which encodes co-evolutionary information contributing to coupling relationship between pairwise residues. Then the *Upside* [15] is launched for

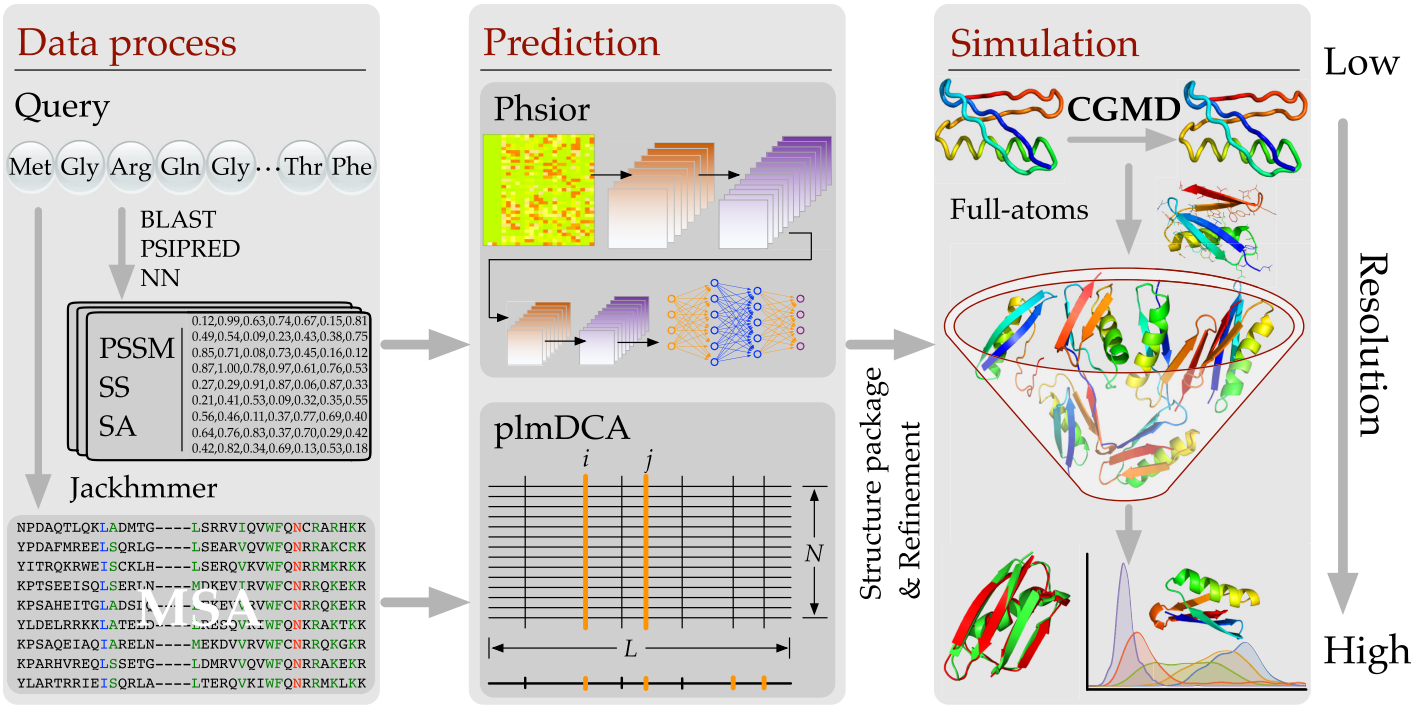

Fig 1. The system flowchart that is used for predicting protein tertiary structure. At the first stage, *NiDelta* constructs both training dataset and MSA for *Phsior* and residue-contacts estimator, respectively. The predicted torsion angles (ϕ, ψ) and estimated residue-contacts are used as restraints for parallelly launching 500 *Upside* simulations, each of which starts with an extended model represented by a simplified structure for sampling its conformation space.

https://doi.org/10.1371/journal.pone.0205819.g001

Table 1. Details of the benchmark proteins and accuracy of predictions achieved by the proposed approach.

Protein name	*L*	Fold	*N*	C_α-RMSD$_{crt}$	C_α-RMSD$_{best}$	Ref. PDB
CrR115	134	α/β	6.0k	4.57 (0.60)	2.51 (0.79)	2lcgA
ER553	141	α/β	98k	4.11 (0.67)	3.11 (0.76)	2k1sA
C-H-RAS P21	166	α/β	574k	4.08 (0.75)	2.98 (0.77)	5p21A
HR2876B	107	α/β	6.9k	4.52 (0.64)	3.42 (0.69)	2ltmA
CG2496	115	α/β	19.8k	2.80 (0.75)	2.19 (0.80)	2kptA
Thioredoxin	105	α/β	214k	2.88 (0.73)	2.12 (0.80)	1rqmA
CheY	130	α/β	887k	8.08 (0.57)	4.21 (0.64)	1e6kA
Ribonuclease HI	143	α/β	63.8k	9.46 (0.42)	5.47 (0.56)	1f21A
Isomerase	108	$\alpha+\beta$	68.4k	5.17 (0.57)	3.34 (0.68)	1r9hA
OR36	134	α/β	6.2k	6.42 (0.47)	4.08 (0.68)	2lciA
MTH1958	136	β	43.9k	7.94 (0.37)	4.77 (0.63)	1tvgA
SgR145	173	α/β	771k	6.87 (0.51)	4.99 (0.63)	3merA
Tpx	167	α/β	185k	3.03 (0.77)	2.38 (0.83)	2jszA
YwIE	150	α/β	40.6k	3.42 (0.76)	2.52 (0.82)	1zggA
FluA	173	β/α	15.9k	7.09 (0.50)	5.02 (0.59)	1n0sA
Rhodopsin II	222	α	3.4k	5.68 (0.64)	5.24 (0.65)	2ksyA
Savinase	269	α/β	102k	6.83 (0.65)	5.17 (0.69)	1svnA
MBP	370	α/β	200k	8.85 (0.51)	6.49 (0.64)	1dmbA

L, Protein length; ***N***, Number of sequences obtained by jackhmmer method; C_α-RMSD$_{crt}$, RMSD in full length of the centroid structure of the largest cluster compared to the native shown in Å(TM-score); C_α-RMSD$_{best}$, RMSD in full length of the best structure compared to the native shown in Å (TM-score).

https://doi.org/10.1371/journal.pone.0205819.t001

protein conformation samplings with the restraints of predicted torsion angles based on convolutional neural network and contacts derived from evolutionary information.

Torsional angles prediction

The benchmark dataset for *Phsior* is collected from RCSB PDB library and pre-culled through PISCES [16]. The library of native protein crystal structures was generated by the PISCES (November 6, 2017) with the following conditions: (1) sequence percentage identity $\leq$ 50%; (2) resolution $\leq$ 1.8Å; (3) R-factor $\leq$ 0.25; (4) sequence length $\geq$ 50. In the dataset, there are 10,586 chains used as the sequence library. The experimental values of the (ϕ, ψ) angles are extracted by STRIDE program [21], and the N- and C-terminal residues are neglected because of the incompleteness of four continuous backbone atoms [22].

Phsior is a real-value predictor developed based on the convolutional neural network for predicting the torsion angles (ϕ, ψ). Briefly, the architecture of *Phsior* is illustrated in Fig 2 (see also S1 Text). *Phsior* extracts three types of sequence-based features involving position-specific scoring matrices (PSSM), secondary structure (SS), and solvent accessibility (SA). The PSSM is generated by PSIBLAST [23] search of the query against a non-redundant sequence database with 20 log-odds scores taken at each position. The secondary structure (SS) is predicted by PSI-PRED [24], with the three states defined as alpha-helix, beta-strand, and coil. The solvent accessibility (SA) is predicted by the neural networks [25]. These three kinds of features will be normalized and used as inputs of the CNN model.

Phsior begins with a simplistic baseline to predict torsion angles (ϕ, ψ) by employing a fixed-size context window of 17 amino acids through two convolutional layers and two fully-connected layers (as illustrated in Fig 2). *Phsior* predicts the torsion angles (ϕ, ψ) of the central amino acid via the final fully-connected layer.

As inputs of the deep network, data is normalized to the range of 0.0 to 1.0. Then we use a window size of 17 to include the neighborhood effect of close amino acids. The data produces a probability map of 35 × 24. The convolutional layers in *Phsior* are to detect recurrent spatial patterns that best represent the local features, while max-pooling layers are to down-sample the features for increasing translational invariance of the network. The fully connected layers are to integrate for the outputs and then make the final predictions for the torsion angles (ϕ, ψ).

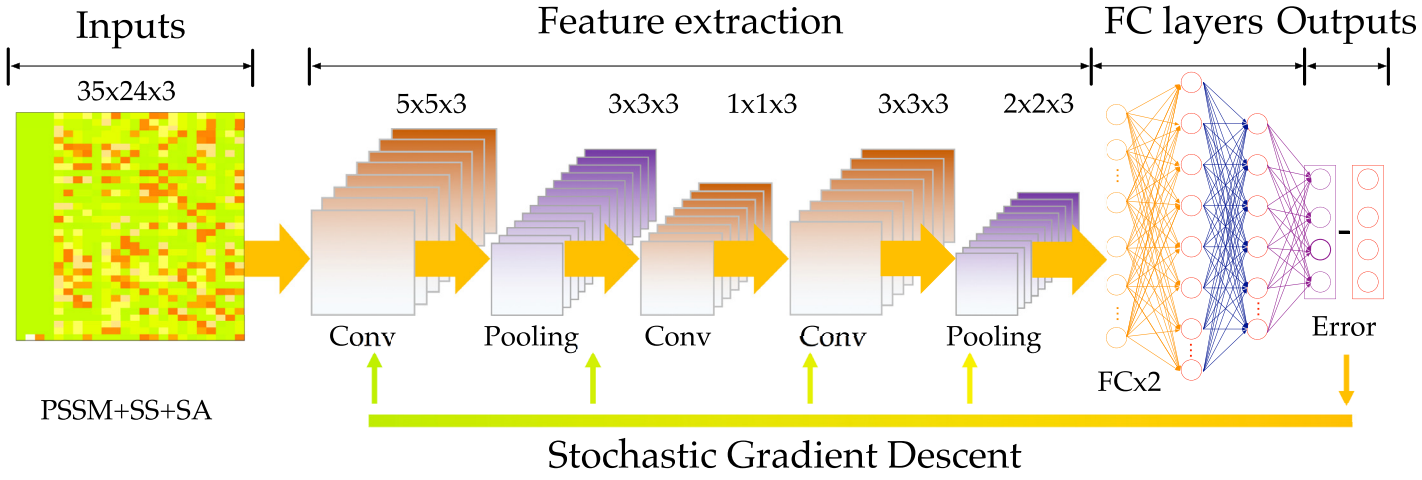

Fig 2. The architecture of *Phsior*. The feature extraction stage includes convolutional and max-pooling layers. The first convolutional layer consists of 16x 5-filters, which slide along the input feature matrix. The second and third convolutional layers work on successive convolutions from previous layers. Following the filters, two fully connected layers are presented to integrate and make final predictions of ϕ and ψ.

https://doi.org/10.1371/journal.pone.0205819.g002

In *Phsior*, a convolutional filter can be interpreted as sliding along the input feature matrix, sharing and/or re-using the same few weights on each local patch of the inputs. Fig 2 illustrates the convolutional layers that work on an example amino acid from training samples. In particular, the first convolutional layer in Fig 2 consists of the 5-filters which is repeated several times as it slides along the feature matrix. Generally, local properties of the input data are important, the small filters show their capability in learning and maintaining information derived from the amino acid sequence at different scales.

In the output layer of *Phsior*, sine and cosine are employed to remove the effect of angle periodicity. Predicted sine and cosine values are converted back to angles by using the equation $\alpha = \tan^{-1}[\sin(\alpha)/\cos(\alpha)]$.

Weights of *Phsior* are randomly initialized according to a zero-centered Gaussian distribution with a standard deviation of $5/\sqrt{N}$ (N is the number of inputs in each layer). Details of each layer in *Phsior* are shown in S1 Table.

Residue contact prediction

Recently, residue-contacts lead *de novo* prediction in a fast progress, like direct coupling analysis (DCA) [26–28], protein sparse inverse covariance (PSICOV) [29] or Gremlin [30, 31] those are all able to disentangle such indirect correlations, and extract direct coevolutionary couplings. These have been found to accurately predict residue-residue contacts—provided a sufficiently large MSA.

Co-evolutionary information encoded in the amino acid sequences highly contributes to residue contacts [26, 27, 29–31]. Accordingly, we estimate pairwise residue contacts from protein multiple sequence alignment (MSA). Firstly, we prepared the MSAs for each studied protein by searching the query sequence against the UniRef100 database [19] using the jackhmmer method [20]. The obtained MSAs were trimmed based on a minimum coverage, which satisfies two basic rules: (1) in the MSA, if the total number of gaps at a single site is more than 50% of the total number of sequences, the site will not be considered in the estimation of residue-contacts; and (2) the percentage of aligned residues between the query and the obtained sequence less than a given threshold ($\leq$ 30% gaps) will be deleted from the MSA.

After filtering the MSA, we start to estimate coupling scores between pairwise residues according to the direct coupling analysis (DCA) algorithm [5, 26, 27, 32]. Given the MSA, we can easily compute the single site frequency $f_i(A_i)$ and joint frequency $f_{ij}(A_i, A_j)$. To maximize the entropy of the observed probabilities, we can calculate the effective pair couplings and single site bias to meet the maximal agreement between the distribution of expected frequencies and the probability model of actually observed frequencies.

$$\begin{cases} P_i(A_i) = \sum\limits_{A_k|k=i} P(A_1, A_2, \cdots, A_L) = f_i(A_i) \\ P_{ij}(A_i, A_j) = \sum\limits_{A_k|k=i,j} P(A_1, A_2, \cdots, A_L) = f_{ij}(A_i, A_j) \end{cases} \tag{1}$$

Maximizing the entropy of the probability model, we can get the statistical model as follows,

$$P(A_1, A_2, \cdots, A_L) = \frac{1}{Z} \exp\left\{ \sum_{i<j} e_{ij}(A_i, A_j) + \sum_i h_i(A_i) \right\}, \tag{2}$$

where Z is a normalization constant, $e_{ij}(\cdot, \cdot)$ is a pairwise coupling, and $h_i(\cdot)$ is a single site bias. The parameters e_{ij} and h_i are estimated by limited-memory BFGS algorithm [33]. Accordingly,

the mathematical definition of the score in pseudo-likelihood maximization Direct-Coupling Analysis (plmDCA) approach [34] is formulated as follows,

$$DI_{ij} = \sum_{A_i,A_j=1}^{m} P_{ij}^{dir}(A_i, A_j) \ln\left(\frac{P_{ij}^{dir}(A_i, A_j)}{f_i(A_i)f_j(A_j)}\right), \tag{3}$$

where DI_{ij} is the direct coupling score between pairwise amino acids at the ith and jth sites in the MSA, and P_{ij}^{dir} is the effective pairwise probability [27]. The top-ranked set of DI_{ij} are converted to contacts between pairwise residues [26, 34].

Ultra-fast molecular dynamics simulation

In the proposed method, we launched a coarse-grained molecular dynamics simulation (CGMD, termed *Upside*) [15] for sampling the conformation space of a given target sequence. In the *Upside*, the model is presented by a reduced chain representation consisting of the backbone N, C_α, and C atoms. The *Upside* launches dynamics simulations of the backbone trace including sufficient structural details (such as side chain structures and free energies). The inclusion of the side chain free energy highly contributes to the smooth the potential governing the dynamics of the backbone trace [15].

In the *Upside*, only the N, C_α, and C atoms for each residue undergo dynamics. An additional term is also added to capture desolvation effects by computing the number of side chains within a hemisphere above the C_β (a derived position from the backbone positions). This simple representation of the protein allows for molecular dynamics much fast on a smooth landscape. The force field in the *Upside* is defined as follows,

$$V = \sum_i V_i^{rama}(\phi_i, \psi_i) + \sum_{i,j \in backbone\&side-chain} V_{ij} + \sum_i v_i^{env}(N_i) \tag{4}$$

where $\sum_i V_i^{rama}(\phi_i, \psi_i)$ is backbone Ramachandran potential from TCB (turn, coil or bridge) Ramachandran probability models in the NDRD backbone library, and V_{ij} is pairwise potential among 5 backbone atoms (C, C_α, N, O, H) and 20 side-chain atoms. And environment term is kinds of solvation energy based on the number of atoms from side-chain and N_i is defined as follows,

$$N_i = \sum_{j,|i-j|>2} \sum_{\chi_i} p(\chi_i) S(|y_i(\chi_i) - y_i^{C_\beta}| - (8\,\text{Å}), (1\,\text{Å})) S(angle(y_i(\chi_i) - y_i^{C_\beta}, d_i^{C_\beta}) + 0.1, 1). \tag{5}$$

In this study, the predicted torsion angles (ϕ, ψ) and the inferred residue contacts are used as restraints to run *Upside* simulations from an extended structure. In the Upside, the pairwise potential used in this study that is sum of two sigmoid functions with Miyazawa-Jernigan (MJ) potential [35] is employed without the multi-position side chains (refer to [15] for more details). The potential function is formulated as

$$V = \frac{e_{in}}{1 + \exp((r - r_{in})/w_{in})} + \frac{e_{out}}{1 + \exp((r - r_{out})/w_{out})}, \tag{6}$$

where, for the side-chain, $e_{in} = 3$, r_{in} is the distance between pairwise amino acids, $w_{in} = 0.2$, e_{out} is MJ energy, $r_{out} = 6.5$, $w_{out} = 0.2$. For the backbone hydrogen bond and backbone-side-chain hydrogen bonds, the settings are: $e_{in} = 6$, $r_{in} = 1.4$, $w_0 = 0.1$, $e_{out} = -4$, $r_{out} = 2.5$, $w_{out} = 0.125$.

For the ith residue, we provide ranges for both ϕ_i and ψ_i, and in this study, we set the ranges as follows: $\phi_i \in [\phi_i^{pred} - 20^\circ, \phi_i^{pred} + 20^\circ]$ and $\psi_i \in [\psi_i^{pred} - 20^\circ, \psi_i^{pred} + 20^\circ]$. This strategy guides the *Upside* sample the Ramachandran map distribution for the secondary structures.

On the other hand, the contacts provide distant restraints for pairwise residues in spacial, which contribute to sample the tertiary structures. According to the design of experiment conducted, we select top 2L residue contacts. The distance of C_β-C_β between pairwise residues that is less than or equals to 7.5Å in the contact potential function makes non-covalent stronger, while it is greater than 7.5Å will make the interaction weaker, as shown in Eq C of S1 Text. For example, if the distance between the pairwise residues are less than or equal to 7.5Å, the Eq C of S1 Text will produce stronger potential energy that reduce the dynamics in protein folding. The *Upside* is configured by setting weights for hydrogen-bond energy, side chain radial scale energy, side chain radial scale inverse radius and side chain radial scale inverse energy to -4.0, 0.2, 0.65 and 3.0, respectively. For each protein sequence, we launched 500 individual simulations starting from the same extended conformation with a duration time of 500,000 and capture conformations at every 500 frames.

Results and discussion

As described in the methods, we sought to provide a template-free prediction system for folding proteins. The approach only depends on sequence information without any structural templates or fragment libraries. We demonstrate the predictive ability of the developed system on a set of candidate structures of proteins over a range of protein size and different folds. The details of eighteen proteins that are collected from the benchmark models of more than 100 residues in refs. [11, 26] are reported in Table 1. According to pre-calculations, each target has less than 50% identity and similarity to each sequence in the training dataset. As illustrated in the table, we present the protein name, PDB id in RCSB database, length of each protein sequence, protein folds, the number of sequences in each MSA, centroid and best C_α-RMSD with corresponding TM-score (computed by TM-score software [36]). All the comparisons of C_α-RMSD and TM-score are computed in full length of each target protein.

We first compare the predictions on the torsion angles (ϕ, ψ) of the target proteins listed in Table 1 among Anglor [22], Spider2 [37], and our model *Phsior* over the eighteen target proteins. For a fair comparison, a criterion is defined by the mean absolute error (MAE) to validate the predicted angles (ϕ, ψ), and the MAE is to measure the average absolute difference between the experimentally determined and predicted angles. Accordingly, the MAE is formulated as follows,

$$MAE = \frac{1}{N}\sum_{i=1}^{N}(P_i - E_i)^2 \tag{7}$$

where N is the number of residues (excluding N- and C-terminals) in a protein. P_i is the predicted value for ith residue, and E_i is the experimental value of jth residue in the protein.

As illustrated in Fig 3 (see also S1 Fig), the proposed *Phsior* and Spider2 [37] are in comparable performances on the target proteins listed in Table 1. They were all better than those of Anglor [22]. The MAE of torsion angle (ϕ, ψ) predicted by Anglor on each protein was almost three times of that of *Phsior* and Spider2, especially on the transmembrane protein Rhodopsin II (PDB ID: 2KSY), the difference remains the largest among all the comparisons. As we know, Anglor is a combined predictor of support vector machine and simple feedforward artificial neural network, while *Phsior* and Spider2 are based on the deep neural network. Accordingly, the better performances could be a result of the powerful capability of the deep learning technique. Although *Phsior* was slightly better than that of Spider2 on several benchmark targets, as shown in Fig 3, *Phsior* is more stable on the predictions.

Since the residues in a region of protein chain are more likely to be related than independent amino acid far away, this 'locality' make the prediction ability of the CNN method more

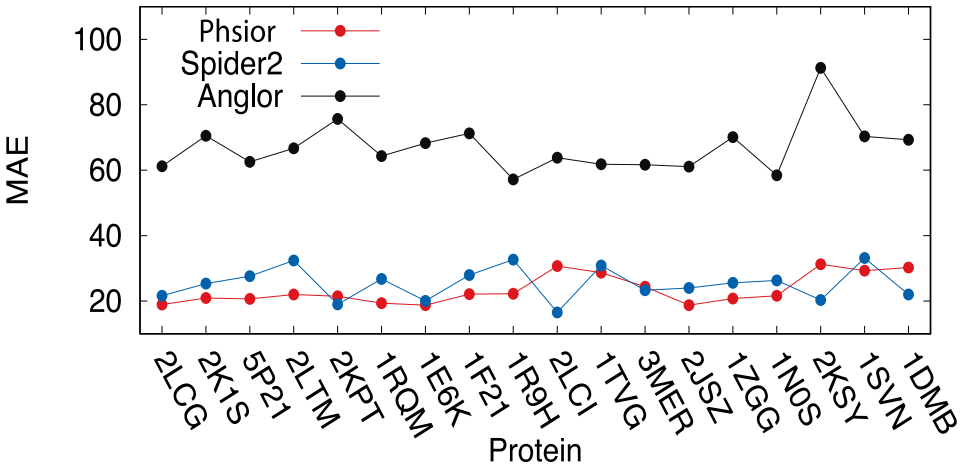

Fig 3. Comparison on the MAE of the predicted torsion angles (ϕ, ψ) among Anglor, Spider2, and *Phsior*.

https://doi.org/10.1371/journal.pone.0205819.g003

powerful. The CNN model can capture the dependences of amino acids in the same chain, which can result in much information of 'locality' among resides. Moreover, the proposed strategy of the predicted torsion angles (ϕ, ψ) can guide the *Upside* to efficiently sample conformation space at high speed. Accordingly, in the developed system, the predictions of *Phsior* are preferred and used as restraints in the *Upside*.

The quality of the predictions by *Phsior* is roughly good to contribute to the restraints for the *Upside* simulation, although there were also several not so good predictions (worse than those of Spider2). However, this did not mean that we could simply use the predicted torsion angles (ϕ, ψ) as starting for the *Upside* simulation. Instead, we found it efficient to pre-defined a range for each torsion angle to launch *Upside* simulations (S1 Text).

We further investigate whether co-evolving sequences can provide sufficient information to specify a good model for assessing blind predictions of protein tertiary structures close to their crystal structures. The predicted residue-contacts mostly correlated with the native ones. As numerous studies [38–40] shown, residue-contacts are significantly important to model the tertiary structure of a protein. The more accurate the predicted residue-contacts are, the better the tertiary model is. In the developed *NiDelta*, these predicted residue-contacts are used as rough restraints to guide and accelerate the molecular dynamics simulation (*Upside*). However, the inferences from the MSA always included noises and false positive predictions, which meant that they could not be simply used for the *Upside*. Instead, we found it efficient and important to generate a potential by sigmoid-like function for the *Upside*. As shown in Eqs (4) and (5) and (C) of S1 Text, the contacts are converted to a potential that makes the *Upside* much robust to the noises in the residue-contacts (see also S1 Text).

For the most of 18 proteins, the estimated residue-contacts include several sparse but informative true positive predictions, making them useful restraints for the *Upside* sampling. Only for the protein OR36 (PDB ID: 2LCI) did NiDelta fail to infer a residue-contact map (S2 Fig), this could result from less diversity in its MAS. Although the bad residue-contacts occur, the *Upside* can be robust to the noises to perform simulation based on Ramachandran map distribution, which could result from the strategy designed in the *NiDelta* for the predicted torsion angles (ϕ, ψ).

As shown in Fig 4, nine representative residue contacts estimated from the MSAs present to compare to the corresponding native ones (see also S2 Fig). The estimated residue-contacts include noises, which (significantly incorrect predictions) are highlighted in green circles in Fig 4. As illustrated, the predicted residue-contacts include numerous noises, that is, many of

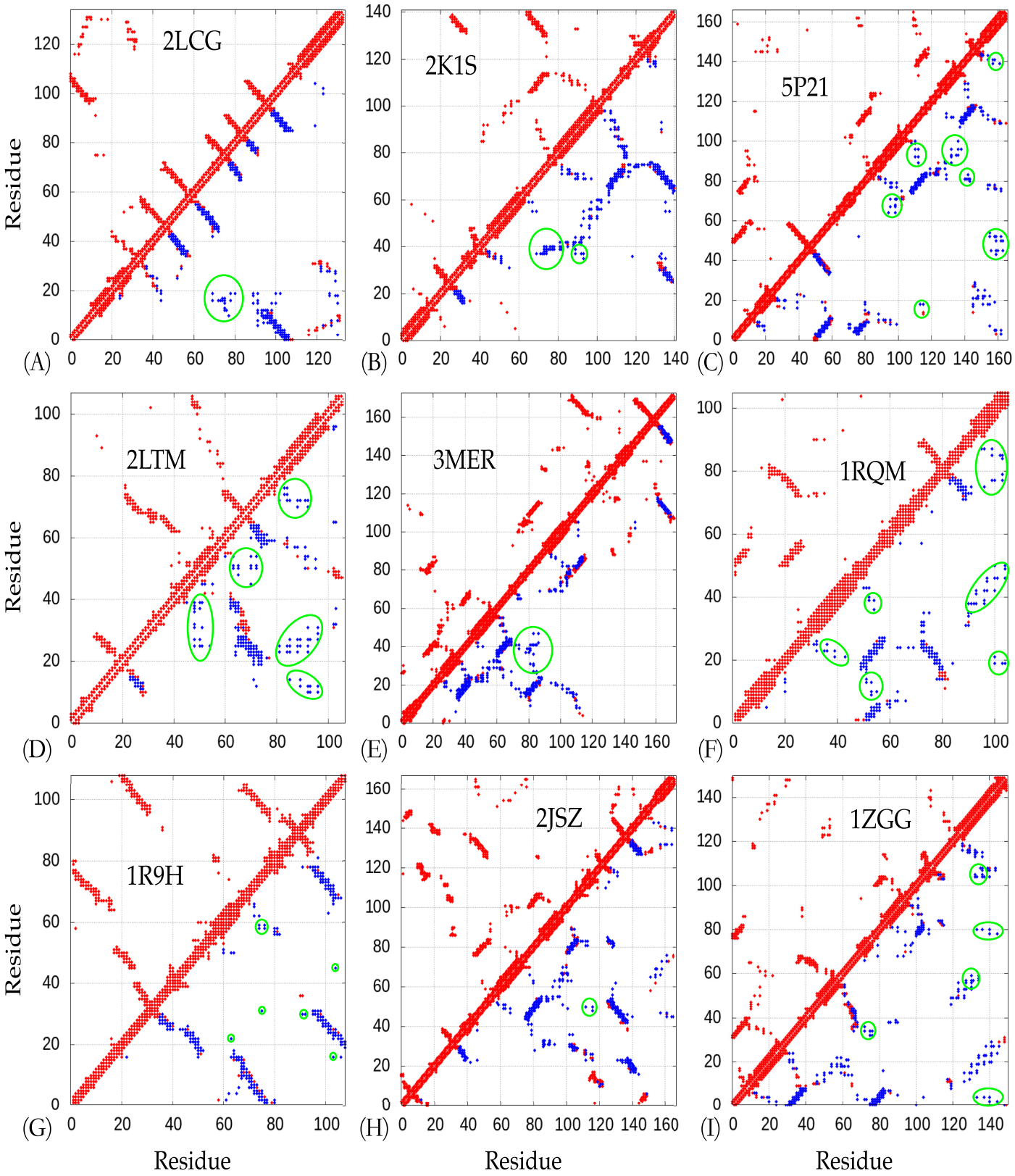

Fig 4. The predicted residue-contacts for highlighted targets. All the residue-contacts (top 2*L*) used in the *Upside* simulations are shown in blue filled squares. The native and estimated residue-contacts are in red and blue, respectively. The dots in green circles are noises (false positive inferences).

https://doi.org/10.1371/journal.pone.0205819.g004

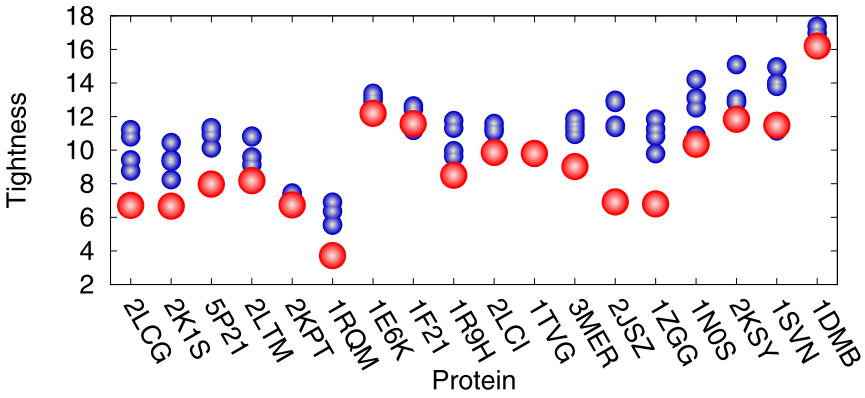

Fig 5. Highlighted predicted structures. Visual comparisons on nine of the target proteins (the native and predicted structures are in red and green, respectively).

https://doi.org/10.1371/journal.pone.0205819.g005

them are incorrect predictions, but the models built by *NiDelta* are not affected so much, which results from the good potentials used in the MD simulation. That is, the developed *NiDelta* is guided by the predicted restraints but not highly dependent on the restraints. For instance, there are five groups of incorrect predictions (noises) in the inferred residue-contacts of the HR2876B protein (PDB ID: 2LTM). The noises possibly led the misfolding of the unstructured regions of the protein as shown in Fig 5. The similarity can also be found in the Thioredoxin (PDB ID: 1RQM) and the YwIE (PDB ID: 1ZGG) proteins.

Immediately after predicting the torsion angles and residue-contacts, it is straightforward to assign the ranges for the angles (ϕ, ψ) and the potentials for interactions between pairwise residues, respectively. Then we launch the ultra-fast coarse-grained molecular dynamics (*Upside* [15]) with the restraints of predicted torsional angles and residue contacts (S1 Text).

For each protein sequence, 500 *Upside* simulations (trajectories) were performed, starting from the unfolded structure. We collected the trajectories for analyzing, and last 50 structures captured from each simulation trajectory were selected from 500 trajectories for clustering (total number is 25,000). As illustrated in S4 Fig, the developed approach can fold a large protein in several CPU hours. We conducted a clustering analysis of the structures using *fast_protein_cluster* software [41] to cluster the structures and calculate the tightness of those clusters, which represent conformational ensembles predicted from each protein sequence. For further study, centroids of the top 5 clusters were selected as our "blind predicted models". The clustering results are illustrated in Fig 6. The biggest cluster has the strongest tightness on the most target proteins (except proteins CG2496, CheY, Ribonuclease HI and Savinase).

To visualize how the structural agreement between the predicted models and the native structure, for nine representative cases, we plotted the proteins corresponding to the best predictions against their C_α-RMSD relative to the experimental reference structures (Fig 5, and see also S3 Fig). The comparison between EVfold [26] and the developed NiDelta on the 18 benchmark proteins as listed in Table 1 is presented in S2 Table. We collected the top 1 predictions from EVfold webserver and the RMSDs and TM-scores of the predictions are illustrated in S2 Table. As illustrated in Fig 5, structural results of the *NiDelta* for nine representative test proteins. In the figure, ribbon models of the lowest C_α – *RMSD* structure (green) (calculated with the *Upside*) superimposed on the corresponding experimental structure (red). For example, as an interesting representative, the C-H-RAS P21 protein p21 (PDB ID: 5P21) involves in a growth promoting signal transduction process [42]. As shown Fig 4(C), although there were noisy predictions in the restraints of torsion angles (ϕ, ψ) (Fig 3 and S1 Fig) and residue-

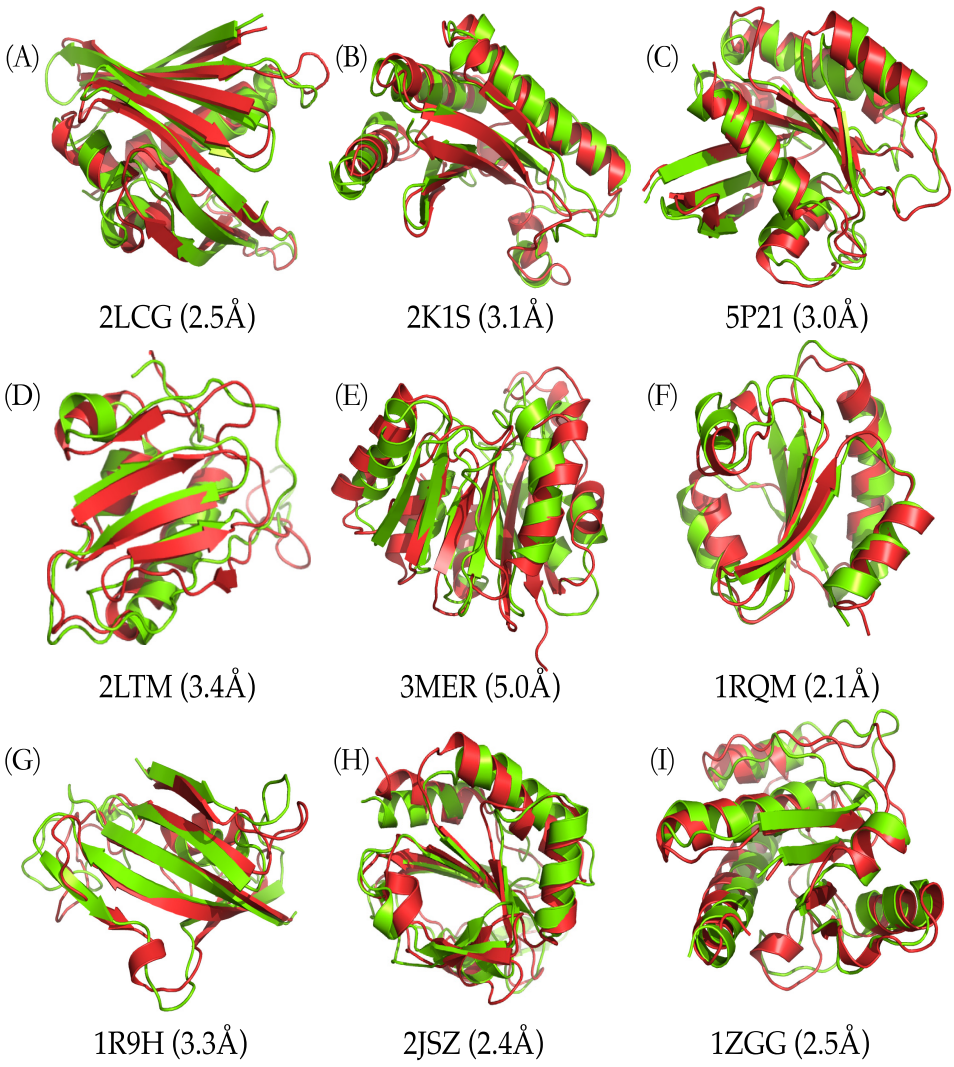

Fig 6. Top five clusters of each target proteins listed in Table 1. The biggest clusters are colored in red, while other clusters are represented in blue.

https://doi.org/10.1371/journal.pone.0205819.g006

residue contacts (Fig 4(C)), The best C_α-RMSD of 3 Å model of the C-H-RAS P21 protein is in the same fold with TM-score of 0.76, and also the centroid model of the largest cluster is blind prediction of C_α-RMSD of 4.1 Å and TM-score of 0.75, which indicates that the *Upside* can be able to fold a large protein and robust to the noises although the existing noises may mislead the simulation in sampling its tertiary structure (e.g. the prediction of the OR36 protein, see Part A in S2 and S3 Figs). As illustrated in Fig 5(F), the structure of the Thioredoxin protein (PDB ID: 1RQM) consists of a central core of a five-stranded β-sheet surrounded by four exposed α-helices [43]. Although the noises and false positive predictions exist in residue contacts (Fig 4), the best C_α-RMSD of the predicted model is 2.1Å, and its corresponding TM-score is as high as 0.8, which mean that the model is almost structurally identity to the native fold. The successful predictions can be also found in the centroid model in top 1 cluster of the C_α-RMSD is 2.9Å and TM-sore 0.73 (Table 1). The blind predictions obtained from the clustering results show that most of the 500 folding simulations converged to similar groups with strength tightness (Fig 6). This could result from that the *Phsior* providing more accurate

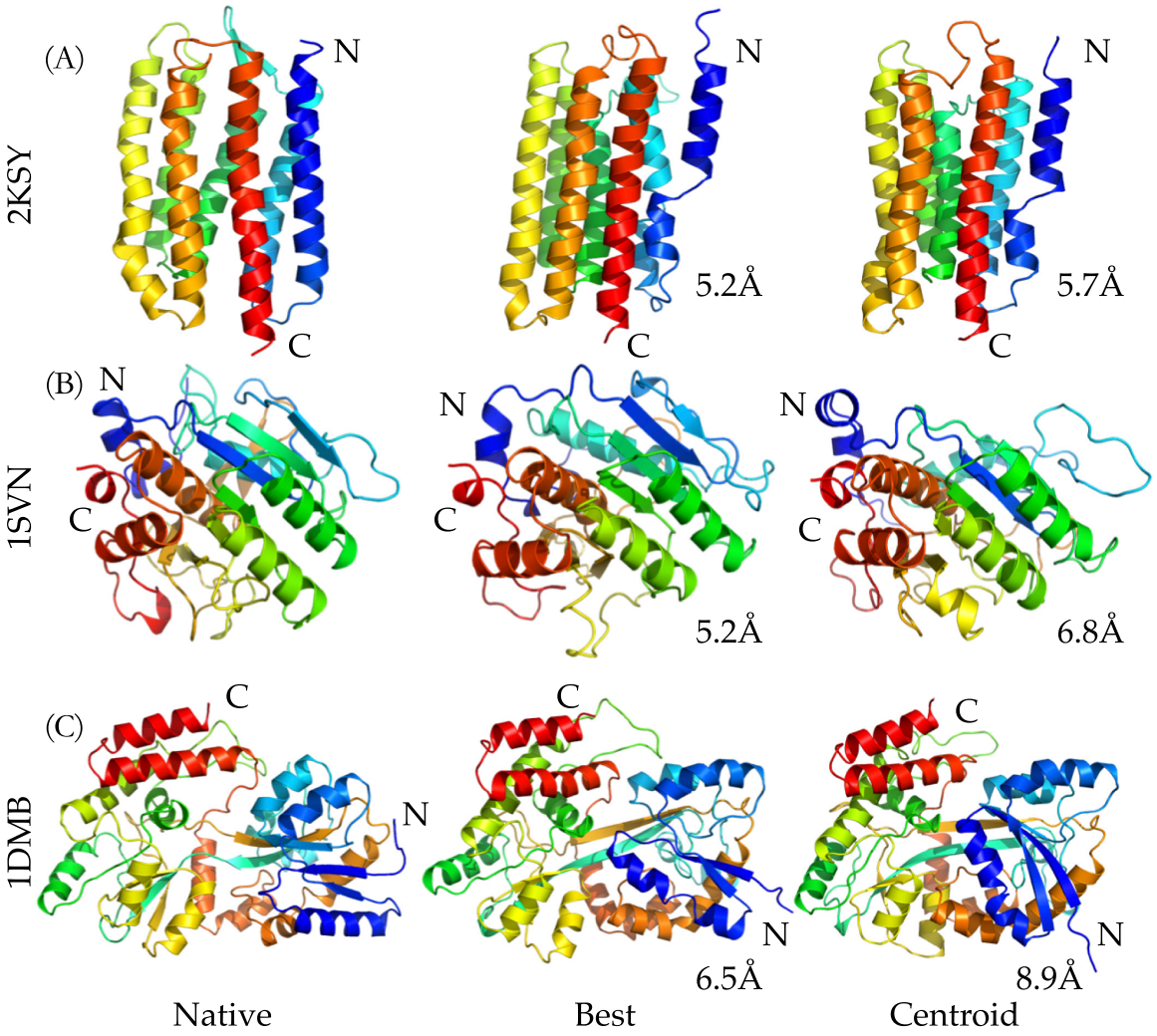

Fig 7. Visual comparisons on three target proteins with more than 200 residues. The highlighted structures from left to right are the native, the structures of the best C_α-RMSD, and the centroid of the biggest cluster, respectively.

https://doi.org/10.1371/journal.pone.0205819.g007

angles (ϕ, ψ) help the *Upside* robust to the noises and inaccurate information. As shown in Fig 5(I) (red), the tertiary fold of the YwlE protein (PDB ID: 1ZGG) is a twisted central four-stranded parallel β-sheet with seven α-helices packing on both sides, in which the active site is favorable for phosphotyrosine binding [44]. The results of the YwlE protein in Figs 4(I), 3 and 6 further demonstrate that *Upside* has a strong predictive ability in folding a protein with inaccurate restraints, even with incorrect information.

Three models (three proteins of more than 200 residues) corresponding to each of the centroid of the biggest clusters are illustrated in Fig 7. The C_α-RMSD values of the centroids compared to the known structures are 5.7Å, 6.8Å, and 8.9Å for Rhodopsin II, Savinase, and MBP proteins, respectively. The protein Rhodopsin II is a membrane protein predicted by the proposed system. For the top ranked predicted model (5.2Å C_α-RMSD with full length alignment, as shown in the center in Fig 7(A)), the terminal helix is misaligned, but the orientations of other six helices are in an excellent agreement with those of the crystal structure. As illustrated in the right of Fig 7(B), the centroid model is also misaligned in the terminal helix, but it provided more structural details as shown in the helices 5 and 6. The structure of the Savinase

protein chosen as the protein of interests has an α/β fold consisting of 9 helices and 9 strands, which is a representative of subtilisin enzymes with maximum stability and high activity [45]. The model of the best C_α-RMSD has correct topography of seven β-strands and eight α-helices, while there are six β-strands and seven α-helices in the centroid model. Flexibility in the conformation occurs in the C-terminal region of Savinase protein [45], which makes the prediction particularly challenging. As shown, both the models of the best C_α-RMSD and centroid capture the structural information. As shown in Fig 7(C), the largest protein tested in the benchmark test is the maltodextrin binding protein (MBP), which is from Escherichia coli serving as the initial receptor for both the active transport of and chemotaxis toward a range of linear maltose sugars [46], with 370 amino acids. It is significantly larger than proteins that can be predicted by other *de novo* computational approaches [26]. With the predicted angles (ϕ, ψ) and residue-contacts, the *Upside* can achieve a blind model of C_α-RMSD 8.9Å and TM-score 0.51, which indicates that the model is in about the same fold [36] and efficiently predictive ability of the proposed approach in the particularly challenging *de novo* structure prediction of large proteins. Accordingly, a strength of the proposed method is demonstrated here is that, based on the centroids of those top 5 clusters, we can potentially develop iterative predictions for larger proteins by collecting centroid models and extracting the informative restraints from previous round of simulations as refinements.

Conclusion

This study presents a way of integrating predicted torsion angles & residue contacts within an ultra-fast molecular dynamics simulation (*Upside*) to achieve *de novo* structure prediction on large proteins. We have tested the proposed approach on the proteins of more than 100 residues and different folds, and also have achieved the agreement of the predictions with the native structures of the benchmark proteins. Statistically determined residue-contacts from the MSAs and torsion angles (ϕ, ψ) predicted by deep learning method provide valuable structural restraints for the ultra-fast MD simulation (*Upside*). The *Upside* provides a simulation with high computational efficiency, which allows users predict structures of large proteins in several CPU hours, get highly accurate models, and details of partial protein folding pathways. Depending on a portion of structural restraints predicted and estimated from the amino acid sequence, the proposed methodology makes the *Upside* a perfect computational platform for *de novo* structure prediction of large proteins.

Although pairwise couplings statistically inferred from protein multiple sequence alignment is a breakthrough in contribution to computational protein structure prediction, there are a number of limitations. For example, residue-residue contacts cannot be estimated if there are no enough as diverse as possible multiple sequences in an alignment of a protein family. Additionally, even when we have sufficient sequences, the pairwise contacts contain false positive predictions that may result in incorrectly building the 3D structure of a protein. Another limitation, applicable to all existing approaches, is predicting the torsion angles (ϕ, ψ). It is challenging to accurately predict torsion angles. *Phsior*, designed based on deep convolutional neural network, is able to predict the angles, but it is difficult to make accurate prediction of each pair (ϕ, ψ). Although we have provided a strategy to handle the inaccurately predicted torsion angles and noised residue-residue contacts, work that of more deep network and iteratively passes information (e.g. averaged torsion angles and contact maps from top 2 structural clusters) collected from previous round of predictions to the next round is currently underway for better predictions of large proteins.

The predicted models (of the best C_α-RMSD and centroid) are consistent with the crystal structures of their natives, and the validation of our approach on eighteen large

proteins suggests that the developed approach is capable in efficiently folding large protein based on predicted restraints. Accordingly, we are confident that future refinement of the approach will be successfully applied to very large proteins and complexes when experimental restraints are available, such as chemical shift, sparse nuclear overhauser effect (NOE) and cryo-electron microscopy (cryo-EM) maps. In summary, we introduce a method *NiDelta* as a *de novo* prediction system for large proteins. We hope this approach will find its place in the fields of both the protein structure prediction and determination in the future.

Supporting information

S1 Text. Supplemental Text.
(PDF)

S1 Fig. Computational time on each protein.
(PDF)

S2 Fig. Comparison on the MAE of the predicted torsion angles (ϕ, ψ) among Anglor, Spider2, and *Phsior*. (a) MAE comparison of ϕ, and (b) MAE comparison of ψ.
(PDF)

S3 Fig. The predicted residue-contacts for highlighted targets listed Table 1. All the residue-contacts (top 2*L*) used in the *Upside* simulations are shown in blue filled squares. The native and estimated residue-contacts are in red and blue, respectively. The dots in green circles are noises (false positive inferences).
(PDF)

S4 Fig. Visual comparisons on the highlighted predicted models of the target proteins.
(PDF)

S1 Table. The key layers in *Phsior* with convolutional and fully connected layers.
(PDF)

S2 Table. The comparison of residue-contact (top L/5 and L/2) among metapsicov, NeBcon and plmDCA.
(PDF)

S3 Table. Comparison (whole length) between EVfold and NiDelta on the benchmark proteins.
(PDF)

Acknowledgments

We thank Drs. T.R. Sosnick, K.F. Freed, J.M. Jumper, and S. Wang for help and advice. This work was supported by DGIST start-up fund No. 2018010089 and the Korean Government Ministry of Trade, Industry and Energy N0001822. We also gratefully acknowledge the DGIST Supercomputing and Big Data Center for dedicated allocation of supercomputing time.

Author Contributions

Conceptualization: Ngaam J. Cheung, Wookyung Yu.

Data curation: Wookyung Yu.

Formal analysis: Ngaam J. Cheung, Wookyung Yu.

Funding acquisition: Wookyung Yu.

Investigation: Ngaam J. Cheung, Wookyung Yu.

Methodology: Ngaam J. Cheung, Wookyung Yu.

Software: Ngaam J. Cheung.

Validation: Ngaam J. Cheung, Wookyung Yu.

Visualization: Ngaam J. Cheung.

Writing – original draft: Ngaam J. Cheung, Wookyung Yu.

Writing – review & editing: Ngaam J. Cheung, Wookyung Yu.

References

1. Röthlisberger D, Khersonsky O, Wollacott AM, Jiang L, DeChancie J, Betker J, et al. Kemp elimination catalysts by computational enzyme design. Nature. 2008; 453(7192):190–195. https://doi.org/10.1038/nature06879 PMID: 18354394
2. Davis IW, Baker D. RosettaLigand docking with full ligand and receptor flexibility. Journal of molecular biology. 2009; 385(2):381–392. https://doi.org/10.1016/j.jmb.2008.11.010 PMID: 19041878
3. Qian B, Ortiz AR, Baker D. Improvement of comparative model accuracy by free-energy optimization along principal components of natural structural variation. Proceedings of the National Academy of Sciences of the United States of America. 2004; 101(43):15346–15351. https://doi.org/10.1073/pnas.0404703101 PMID: 15492216
4. Ovchinnikov S, Park H, Varghese N, Huang PS, Pavlopoulos GA, Kim DE, et al. Protein structure determination using metagenome sequence data. Science. 2017; 355(6322):294–298. https://doi.org/10.1126/science.aah4043 PMID: 28104891
5. Marks DS, Hopf TA, Sander C. Protein structure prediction from sequence variation. Nature Biotechnology. 2012; 30:1072–1080. https://doi.org/10.1038/nbt.2419 PMID: 23138306
6. Anfinsen CB. The formation and stabilization of protein structure. Biochemical Journal. 1972; 128 (4):737–749. https://doi.org/10.1042/bj1280737 PMID: 4565129
7. Šali A, Blundell TL. Comparative Protein Modelling by Satisfaction of Spatial Restraints. Journal of Molecular Biology. 1993; 234(3):779–815. https://doi.org/10.1006/jmbi.1993.1626 PMID: 8254673
8. Kinch LN, Li W, Monastyrskyy B, Kryshtafovych A, Grishin NV. Evaluation of free modeling targets in CASP11 and ROLL. Proteins: Structure, Function, and Bioinformatics. 2016; 84(S1):51–66. https://doi.org/10.1002/prot.24973
9. Zhang W, Yang J, He B, Walker SE, Zhang H, Govindarajoo B, et al. Integration of QUARK and I-TASSER for *Ab Initio* Protein Structure Prediction in CASP11. Proteins: Structure, Function, and Bioinformatics. 2016; 84(S1):76–86. https://doi.org/10.1002/prot.24930
10. Das R, Baker D. Macromolecular modeling with Rosetta. Annu Rev Biochem. 2008; 77:363–382. https://doi.org/10.1146/annurev.biochem.77.062906.171838 PMID: 18410248
11. Shen Y, Bax A. Homology modeling of larger proteins guided by chemical shifts. Nature methods. 2015; 12(8):747–750. https://doi.org/10.1038/nmeth.3437 PMID: 26053889
12. Bradley P, Misura KM, Baker D. Toward high-resolution *de novo* structure prediction for small proteins. Science. 2005; 309(5742):1868–1871. https://doi.org/10.1126/science.1113801 PMID: 16166519
13. Kim DE, Blum B, Bradley P, Baker D. Sampling bottlenecks in *de novo* protein structure prediction. Journal of molecular biology. 2009; 393(1):249–260. https://doi.org/10.1016/j.jmb.2009.07.063 PMID: 19646450
14. Söding J. Big-data approaches to protein structure prediction. Science. 2017; 355(6322):248–249. https://doi.org/10.1126/science.aal4512 PMID: 28104854
15. Jumper JM, Freed KF, Sosnick TR. Maximum-likelihood, self-consistent side chain free energies with applications to protein molecular dynamics. arXiv preprint arXiv:161007277. 2016;.
16. Wang G, Dunbrack RL Jr. PISCES: a protein sequence culling server. Bioinformatics. 2003; 19 (12):1589–1591. https://doi.org/10.1093/bioinformatics/btg224 PMID: 12912846
17. LeCun Y, Haffner P, Bottou L, Bengio Y. Object recognition with gradient-based learning. Shape, contour and grouping in computer vision. 1999; p. 823–823.
18. Sibe web-server;. Available from: http://wyu.dgist.ac.kr/sibe/feature.html [cited 15.09.2017].

19. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, Consortium U. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics. 2015; 31(6):926–932. https://doi.org/10.1093/bioinformatics/btu739 PMID: 25398609

20. Eddy SR. Accelerated Profile HMM Searches. PLOS Computational Biology. 2011; 7(10):1–16. https://doi.org/10.1371/journal.pcbi.1002195

21. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins: structure, function, and genetics. 1995; 23(4):566–579. https://doi.org/10.1002/prot.340230412

22. Wu S, Zhang Y. ANGLOR: A Composite Machine-Learning Algorithm for Protein Backbone Torsion Angle Prediction. PLoS ONE. 2008; 3(10):e3400. https://doi.org/10.1371/journal.pone.0003400 PMID: 18923703

23. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research. 1997; 25 (17):3389–3402. https://doi.org/10.1093/nar/25.17.3389 PMID: 9254694

24. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. Journal of molecular biology. 1999; 292(2):195–202. https://doi.org/10.1006/jmbi.1999.3091 PMID: 10493868

25. Chen H, Zhou HX. Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. Nucleic acids research. 2005; 33(10):3193–3199. https://doi.org/10.1093/nar/gki633 PMID: 15937195

26. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, et al. Protein 3D Structure Computed from Evolutionary Sequence Variation. PLOS ONE. 2011; 6(12):1–20. https://doi.org/10.1371/journal.pone.0028766

27. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences. 2011; 108(49):E1293–E1301. https://doi.org/10.1073/pnas.1111471108

28. Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. Physical Review E. 2013; 87(1):012707. https://doi.org/10.1103/PhysRevE.87.012707

29. Jones DT, Buchan DW, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics. 2011; 28 (2):184–190. https://doi.org/10.1093/bioinformatics/btr638 PMID: 22101153

30. Balakrishnan S, Kamisetty H, Carbonell JG, Lee SI, Langmead CJ. Learning generative models for protein fold families. Proteins: Structure, Function, and Bioinformatics. 2011; 79(4):1061–1078. https://doi.org/10.1002/prot.22934

31. Kamisetty H, Ovchinnikov S, Baker D. Assessing the utility of coevolution-based residue–residue contact predictions in a sequence-and structure-rich era. Proceedings of the National Academy of Sciences. 2013; 110(39):15674–15679. https://doi.org/10.1073/pnas.1314045110

32. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. Identification of direct residue contacts in protein-protein interaction by message passing. Proceedings of the National Academy of Sciences. 2009; 106 (1):67–72. https://doi.org/10.1073/pnas.0805923106

33. Nocedal J. Updating quasi-Newton matrices with limited storage. Mathematics of computation. 1980; 35(151):773–782. https://doi.org/10.1090/S0025-5718-1980-0572855-7

34. Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. Physical Review E. 2013; 87(1):012707. https://doi.org/10.1103/PhysRevE.87.012707

35. Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules. 1985; 18(3):534–552. https://doi.org/10.1021/ma00145a039

36. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins: Structure, Function, and Bioinformatics. 2004; 57(4):702–710. https://doi.org/10.1002/prot.20264

37. Heffernan R, Paliwal K, Lyons J, Dehzangi A, Sharma A, Wang J, et al. Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. Scientific reports. 2015; 5:11476. https://doi.org/10.1038/srep11476 PMID: 26098304

38. Jones DT, Singh T, Kosciolek T, Tetchner S. MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. Bioinformatics. 2014; 31 (7):999–1006. https://doi.org/10.1093/bioinformatics/btu791 PMID: 25431331

39. He B, Mortuza S, Wang Y, Shen HB, Zhang Y. NeBcon: protein contact map prediction using neural network training coupled with naïve Bayes classifiers. Bioinformatics. 2017; 33(15):2296–2306. https://doi.org/10.1093/bioinformatics/btx164 PMID: 28369334

40. Schaarschmidt J, Monastyrskyy B, Kryshtafovych A, Bonvin AM. Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age. Proteins: Structure, Function, and Bioinformatics. 2018; 86:51–66. https://doi.org/10.1002/prot.25407

41. Hung LH, Samudrala R. fast_protein_cluster: parallel and optimized clustering of large-scale protein modeling data. Bioinformatics. 2014; 30(12):1774–1776. https://doi.org/10.1093/bioinformatics/btu098 PMID: 24532722

42. Barbacid M. Ras genes. Annual Review of Biochemistry. 1987; 56(1):779–827. https://doi.org/10.1146/annurev.bi.56.070187.004023 PMID: 3304147

43. Leone M, Di Lello P, Ohlenschläger O, Pedone EM, Bartolucci S, Rossi M, et al. Solution structure and backbone dynamics of the K18G/R82E Alicyclobacillus acidocaldarius thioredoxin mutant: a molecular analysis of its reduced thermal stability. Biochemistry. 2004; 43(20):6043–6058. https://doi.org/10.1021/bi036261d PMID: 15147188

44. Xu H, Xia B, Jin C. Solution structure of a low-molecular-weight protein tyrosine phosphatase from Bacillus subtilis. Journal of bacteriology. 2006; 188(4):1509–1517. https://doi.org/10.1128/JB.188.4.1509-1517.2006 PMID: 16452434

45. Betzel C, Klupsch S, Papendorf G, Hastrup S, Branner S, Wilson KS. Crystal structure of the alkaline proteinase Savinase™ from Bacillus lentus at 1.4 Å resolution. Journal of molecular biology. 1992; 223 (2):427–445. https://doi.org/10.1016/0022-2836(92)90662-4 PMID: 1738156

46. Sharff AJ, Rodseth LE, Quiocho FA. Refined 1.8-. ANG. structure reveals the mode of binding of. beta.-cyclodextrin to the maltodextrin binding protein. Biochemistry. 1993; 32(40):10553–10559.