Computational and Structural Biotechnology Journal 19 (2021) 2575–2587

ELSEVIER

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Multiobjective heuristic algorithm for *de novo* protein design in a quantified continuous sequence space

Rui-Xiang Li [a,1], Ning-Ning Zhang [c,1], Bin Wu [d], Bo OuYang [c,*], Hong-Bin Shen [a,b,*]

[a] *Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China*

[b] *Department of Computer Science, Shanghai Jiao Tong University, Shanghai 200240, China*

[c] *State Key Laboratory of Molecular Biology, CAS Center for Excellence in Molecular Cell Science, Shanghai Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Shanghai 201203, China*

[d] *National Facility for Protein Science in Shanghai, ZhangJiang Lab, Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai 201210, China*

ARTICLE INFO

Article history:
Received 17 December 2020
Received in revised form 19 April 2021
Accepted 22 April 2021
Available online 25 April 2021

Keywords:
Protein design
Bioinformatics
Optimization

ABSTRACT

Protein design usually involves sequence search process and evaluation criteria. Commonly used methods primarily implement the Monte Carlo or simulated annealing algorithm with a single-energy function to obtain ideal solutions, which is often highly time-consuming and limited by the accuracy of the energy function. In this report, we introduce a multiobjective algorithm named Hydra for protein design, which employs two different energy functions to optimize solutions simultaneously and makes use of the latent quantitative relationship between different amino acid types to facilitate the search process. The framework uses two kinds of prior information to transform the original disordered discrete sequence space into a relatively ordered space, and decoy sequences are searched in this ordered space through a multiobjective swarm intelligence algorithm. This algorithm features high accuracy and a high-speed search process. Our method was tested on 40 targets covering different fold classes, which were computationally verified to be well folded, and it experimentally solved the 1UBQ fold by NMR in excellent agreement with the native structure with a backbone RMSD deviation of 1.074 Å. The Hydra software package can be downloaded from: http://www.csbio.sjtu.edu.cn/bioinf/HYDRA/ for academic use.

1. Introduction

Proteins play a critical role in biological activities, serving as, for instance, energy producers, structural bricks, sensors, and catalysts [1]. In recent years, *de novo* protein design has become an emerging technology with a significant impact on the development of medicine, nanoscience, catalytic chemistry and other fields. Computational protein design (CPD) is used to calculate the appropriate protein sequence given the target protein scaffold by computer algorithms, and it can be regarded as the inverse procedure of protein structure prediction [2]. For fixed backbone protein design, there are at least two fundamental questions: the sequence searching algorithm in the large sequence space; sequence evaluation methods or energy function. Large deviations of the designed sequences and long search processes in the disordered discrete sequence space are among the primary issues facing current protein design methods.

Recently, several different protein design methods have been developed, most of which have used the Replica-Exchange Monte Carlo (REMC) or simulated annealing (SA) algorithms minimizing a single energy function, such as Rosetta [3], SEF_V [4], and etc. These methods use many iterations (more than 10^5) to obtain an ideal solution. How to evaluate the designed sequence is an important issue, which is usually formed as an energy function in the optimization protocol. For instance, in the protein structure prediction area, many energy functions have been proposed and widely used, i.e., the CHARMM [5], Rosetta [6], DFIRE [7], UNRES [8], and Amber [9], and etc. Some of them are derived from the point of physics atomic interactions, and some are from the statistical knowledge-based descriptions. Previous studies have shown that there is not a general better energy function for all proteins and

[*] Corresponding authors at: Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, 800 Dongchuan Road, China (H.-B. Shen); Shanghai Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, 333 Haike Road, Shanghai 201203, China (B. OuYang).

E-mail addresses: ouyang@sibcb.ac.cn (B. OuYang), hbshen@sjtu.edu.cn (H.-B. Shen).

[1] Co-first author

https://doi.org/10.1016/j.csbj.2021.04.046
2001-0370/

how to choose an appropriate energy function has become a difficult problem.

In current protein design area, most programs have applied a single energy function to evaluate the designed sequences, which could also face the same difficulty of the applicability on different proteins. Although many versions of energy functions have been published, it is still difficult to determine which one is the best. It is reasonable for a multiobjective evaluation to combine the advantages of different energy functions, potentially increasing the performance of each iteration. In recent years, some heuristic multi-objective protein design algorithms have been reported and show promising results [10]. One of the major merits of the multi-objective optimization-based protein design is we can evaluate the searched potential sequences from different point of view through the parallelly optimized objectives or energies, which will result in the so-called non-dominated solutions in the high-dimensional space composed by the applied energy functions. This is quite different from optimizing a single energy function in the 1D space, which is usually composed by weighed sum of multiple terms.

In terms of searching algorithms implemented in the original disordered discrete sequence space, it would be difficult to take advantage of the latent quantitative relationship of different amino acid types to simplify the search process. Since the chemical properties of amino acids are relative, their chemical relationship can be used as a guideline in the searching process between iterations, as well as to minimize the searching space in each iteration. Therefore, instead of searching in the original sequence space, it would be better to transform the original discrete sequence space into an ordered continuous space. The REMC or SA algorithm is generally difficult for implementation in continuous space, since the random mutation operation is difficult to guide according to the previous iteration. Searching in continuous space would decrease the total iterations dramatically. Thus, we can use fewer iterations to complete the whole searching process, and multiple energy functions or more complex evaluation methods can be implemented into each iteration.

Thus, we developed *de novo* a new computational fixed backbone protein design method named Hydra. In Hydra, we used the fused prior information collected from the PDB and DSSP databases to transform the original discrete sequence space into a continuous space. A particle swarm optimization [11] (PSO)-based biobjective (double objective functions) swarm intelligence algorithm was implemented in the continuous space to search sequences. Two different energy functions, the FoldX energy function [12] and a function describing local structural properties similar to EvoDesign [13], were used to evaluate the decoy sequences.

We demonstrated that two energy functions could restrain and complement each other to increase the robustness of the selected solutions, and the space transformation could facilitate the search process. We compared the performance of Hydra with two established methods, Rosetta fixed backbone design [3] and ABACUS [4], and the results suggested that the multiobjective optimization algorithm was superior to single objective optimization in protein design, while the space transformation procedure can shorten the search process. The sequences designed by Hydra could be well folded with high accuracy and fewer iterations, as confirmed by the experimentally solved structure of target 1UBQ.

2. Methods

The Hydra tries to solve the fixed backbone protein design problem, so the choice of conformational space is not needed to be considered here. Its overall procedure consists of four steps: construction of prior information, sequence space transformation, iterative biobjective optimization and selection of output sequences. The outline of Hydra is shown in Fig. 1a. We first collected the prior information (fused by the structural and statistical information) from two databases and then transformed the original disordered discrete sequence space into an ordered continuous space (Fig. 1b), where algorithm optimization was implemented. After a certain number of such iterations, several Pareto solutions would be stored in the Pareto archive set. Finally, reliable solutions would be selected from the Pareto archive set as output sequences.

2.1. Structure and statistical information construction

Inspired by EvoDesign [14], the input target scaffold's structural information was obtained from a nonredundant set of the PDB library by the protein structure alignment program TM-align [15], and the target scaffold was used as the probe in the alignment procedure. For each alignment, the TM-align program will return a TM-score [16] value to assess the structure similarity of one protein from the PDB library to the target protein. The TM-score ranges from 0 to 1, with a higher value indicating a higher structural similarity, and a TM-score > 0.5 roughly means that two structures belong to the same fold according to the database analysis [17]. In our method, the PDB database was built from the RCSB PDB database [18] on the basis of several filter conditions:

[1] Proteins are from *Homo sapiens*, [2] candidate structures are determined by X-ray diffraction, [3] resolution is better than 2.5 Å, [4] sequence identity cut-off is 40%. The target scaffold was aligned with all the protein structures in the PDB database, and they were ranked according to their TM-score values relative to the target from highest to lowest.

In general, we collected all the proteins in database with a TM-score > 0.7 to extract the information for the target's homologous structural family. If the number of proteins with a TM-score > 0.7 was less than ten, the cut-off value would gradually decrease to ensure sufficient analogues of the target scaffold. The sequences of the collected proteins will be used to construct a $L \times 20$ position-specific scoring matrix by multiple sequence alignment [19,20] (MSA), where L is the number of input target residues and 20 is the number of amino acid types. This matrix specifies the distribution of different amino acids at each residue position of the target scaffold, which indicates the conservation and variation information for the target's homologous protein family. For residue r at residue position p, the element of this matrix can be formulated as $M(p,r) = \sum_{x=1}^{20} f(p,x) \times B(x,r)$. In this equation, $f(p,x)$ is residue x's appearance frequency at position p in the MSA, and $B(x,r)$ is the BLOSUM62 [21] matrix value of residue pair r and x. It has been demonstrated that the structural information obtained from the target scaffold's homologous protein family is more accurate and robust than the conventional homologous protein information deduced from sequence-based searches, such as PSI-BLAST [22,23].

The statistical information was obtained from the distribution of all the different kinds of amino acid types given certain structural properties of each residue position [4]. This distribution was derived from a DSSP database built with the DSSP files assigned from the PDB database by DSSP standalone software [24], except that some incomplete proteins were excluded. A statistical information matrix will be generated, and one element of this matrix $S(r,p)$ means the value of amino acid r at position p. $S(r,p)$ is determined by the probability distribution of amino acid type r conditioned on the structural properties associated with the corresponding position along the target scaffold. It can be represented as follows:

$$S(r,p) = -\ln p(r|\textit{structure properties at position } p) \quad (1)$$

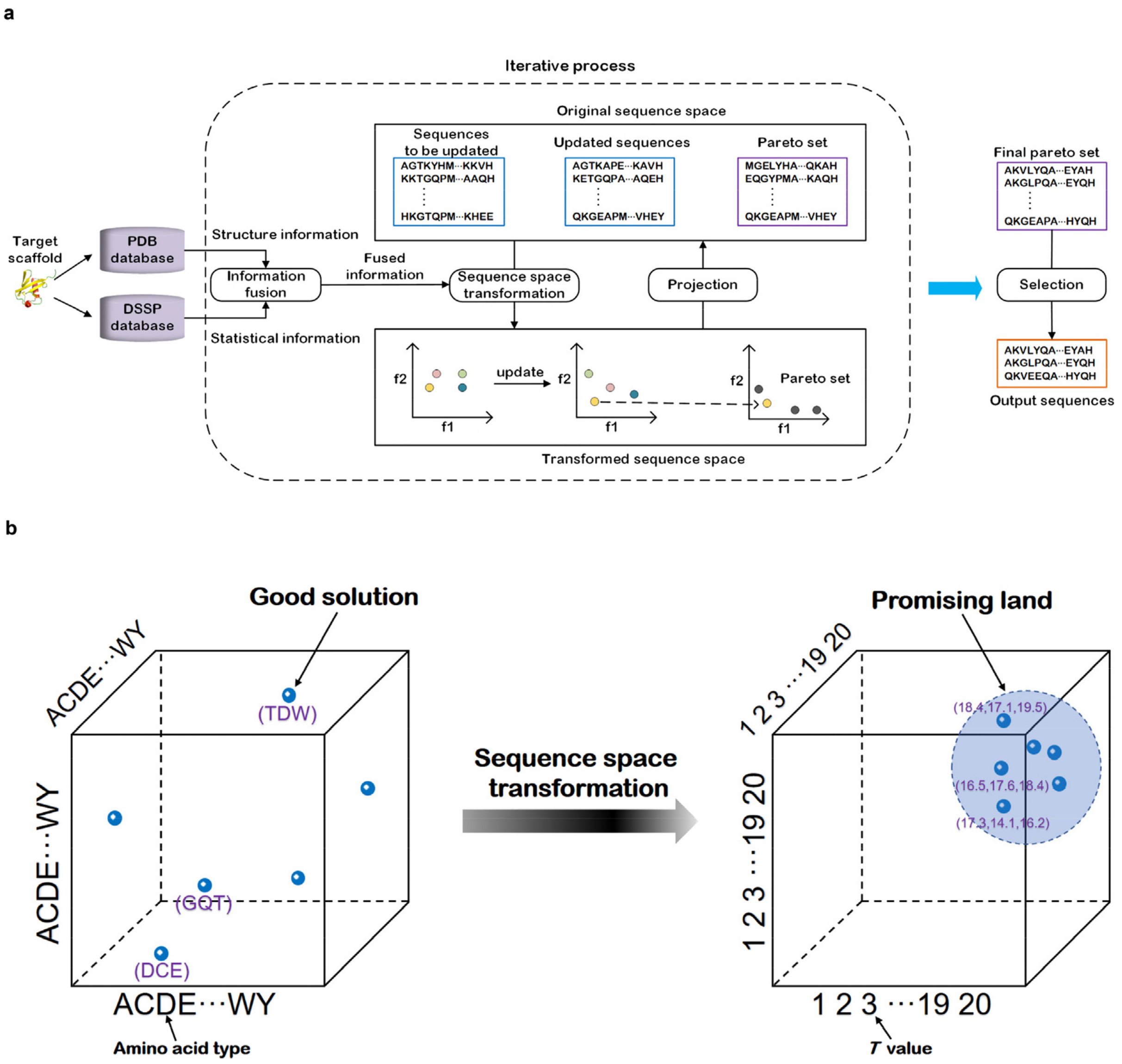


Fig. 1. Flowchart of the Hydra Design, where f1 and f2 are the two energy functions of this algorithm. **a**, Flowchart of Hydra. **b**, Diagram of the space transformation. The left side is the original discrete sequence space. The right side is the quantified continuous space. After space transformation, the ideal solutions (blue orbs) will be clustered in a small area. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Each scaffold position's structural properties include backbone torsional angles, secondary structure (SS) and solvent accessibility surface area (SASA). These structural properties should be met simultaneously when counting the numbers of different amino acid types in the DSSP database given a residue position of the target, and the conditional probability distribution in Eq. (1) was calculated from these counted numbers. To obtain the distribution of different amino acid types associated with certain structural properties, every target residue position was mapped to a target point into an abstract space spanned by different structural properties, and the same procedure was performed for the training proteins in the DSSP database. Neighboring training data points relative to the target data point within the threshold value would be collected. The threshold value was a key point in the accuracy of the statistical information. A higher threshold value would generate more data samples, which could make the statistics more robust and stable. Additionally, the probability distribution was more accurate with a lower threshold value, but the statistical data samples might not be sufficient to construct a reliable probability distribution in such a situation. In this paper, we set the threshold value of the torsional angle difference and solvent accessibility surface area difference as 15 and 20 Å, respectively, according to the empirical data.

2.2. Sequence space transformation

The original sequence space was an L dimensional discrete space, where L was the target protein's sequence length, and there were 20 amino acid types in each dimension. No numerical relationship appeared to exist between these amino acid types, which indicated that they could not be compared in a quantitative manner. Thus, the original sequence space was a high-dimensional and disordered discrete space. In prevalent protein design methods, the sequence search process is directly conducted in the original sequence space [25], which makes it a complex and difficult problem. In our method, the search process will be conducted in a transformed space.

Before transformation, we should normalize the structure information matrix and statistical information matrix obtained from the previous step as follows:

$$M(r,p)' = \frac{M(r,p) - \min_r M(r,p)}{\max_r M(r,p) - \min_r M(r,p)},\ S(r,p)' = \frac{S(r,p) - \min_r S(r,p)}{\max_r S(r,p) - \min_r S(r,p)} \tag{2}$$

With the normalized information matrix, we obtain

$$F(r,p) = \lambda_1 M(r,p)' + \lambda_2 S(r,p)' \tag{3}$$

$M(r,p)'$ and $S(r,p)'$ respectively are elements of the structure information matrix and statistical information matrix, and λ_1 and λ_2 are the coefficients to control the contribution weight of the two information matrices. In our method, the coefficients will vary at regular intervals, which could oscillate the sequence space gradually during the search process. Generally, we do not know the absolute accurate quantitative relationship between the twenty different amino acids, which means that the most ideal distribution of good solutions may not be achieved in the transformed space. Thus, this oscillation will allow more opportunities to discover the desired protein sequences and prevent the search process from stagnating in local minima.

In Eq. (3), $F(r,p)$ is amino acid r's fusion value at sequence position p. In general, a larger fusion value means a greater possibility of being the appropriate amino acid type for its residue position of the target. Thus, in each residue position, we obtain a quantified relationship between the twenty amino acid types. As a result, the original discrete and disordered sequence space is transformed into a continuous and ordered sequence space. For each residue position of the target scaffold, the twenty different amino acids will be sorted by their fusion values. Then, in the transformed sequence space, each amino acid will be assigned a transformed value (T value) according to its fusion value rank, e.g., the amino acid with the largest fusion value will be assigned a value of 20, and the smallest fusion value will be denoted as 1. The transformed sequence space can be written as:

$$R^n = R \times R \cdots \times R = \{(x_1, x_2, \cdots, x_n) | x_k \in R, k = 1, 2, \cdots, n\} \tag{4}$$

n is the target protein's sequence length. The transformed value is:

$$T(r,p) = rank(F(r,p)),\ rank \in \{1, 2, \cdots, 20\} \tag{5}$$

As mentioned previously, the sorting of the fusion value is from small to large. Thus,

$$\begin{cases} T(r_k,p) = rank(F(r_k,p)) = 1, & when\, F(r_k,p) = \min(F(r,p)) \\ T(r_k,p) = rank(F(r_k,p)) = 20, & when\, F(r_k,p) = \max(F(r,p)) \end{cases} \tag{6}$$

2.3. Biobjective directional sequence space search

A. Energy functions

In this paper, the sequence search process was guided by two energy functions, making it a biobjective optimization problem. Two energy functions were used to evaluate the fitness of the decoy sequences with different criteria, and both could play important roles in the structural packing of decoy sequences.

The first energy function was a physics-based force field from FoldX [12], which provided a rapid and quantitative estimation of the interactions that contributed to the stability of proteins. The predictive power of FoldX has been tested on a very large scale of protein structures [12]. FoldX consists of many different force field terms that can be written as follows:

$$f_1 = w_1 E_{vdw} + w_2 E_{solvH} + w_3 E_{solvP} + E_{wb} + E_{hbond} + E_{el} + E_{Kon} + w_4 E_{Smc} + w_5 E_{Ssc} \tag{7}$$

where E_{vdw} is the sum of the van der Waals contributions of all atoms; E_{solvH} and E_{solvP} are the solvation energy for apolar and polar groups; E_{wb} is the extra stabilizing free energy provided by a water molecule; E_{hbond} is the free energy difference between the formation of an intramolecular hydrogen bond compared with intermolecular hydrogen bond formation; E_{el} accounts for the electrostatic contribution of charged groups; E_{Kon} reflects the effect of electrostatic interactions on the association constant kon; and E_{Smc} and E_{Ssc} are deduced from the theoretical estimate. The detailed description of the FoldX force field can be found in the corresponding papers [12]. We used the default parameter setting for the calculation in our method. Since FoldX is a full-atomic based software, to calculate the FoldX value, we should build the corresponding side chain atoms to the target scaffold according to the candidate sequence. In this study, we used SCWRL4.0 [26], a user-friendly protein side chain conformation building software with its own rotamer library, to reconstruct the input target scaffold according to the generated decoy sequences before FoldX calculation. Thus, the protein sequences could be evaluated quantitatively to estimate whether they were appropriate sequences for the target structure.

The second energy function was based on local structural properties inspired from EvoDesign [13]. It is composed of several terms as follows:

$$f_2 = \sum w_1 \Delta SS(p) + w_2 \Delta SA(p) + w_3 (\Delta\phi(p) + \Delta\psi(p)) \tag{8}$$

where $\Delta SS(p)$ is the secondary structure difference between the target scaffold and decoy sequence at position p, which can be formulated as:

$$\Delta SS = \begin{cases} 1, & different\, SS\, type \\ 0, & identical\, SS\, type \end{cases} \tag{9}$$

ΔSA is the difference in solvent accessibility, and $\Delta\phi(p)$ and $\Delta\psi(p)$ are the differences of torsional angles. These three terms are all normalized by:

$$norm(\Delta x) = \frac{\Delta x - \Delta x_{\min}}{\Delta x_{\max} - \Delta x_{\min}} \tag{10}$$

where Δx_{max} and Δx_{min} respectively are the maximum and minimum value of Δx. The decoy sequence's local structural properties were obtained through prediction in each iteration of our algorithm, applying the SPIDER3-Single [27] to fulfil this task, which is a fast secondary structure prediction software based on a single sequence only. It can consistently achieve Q3 accuracy of 72.5% with considerably less running time. The structural properties of the target scaffold were assigned by the DSSP program. The weights ω_1, ω_2 and ω_3 are decided by the relative accuracy of the individual feature predictions, ie. $\omega_1 = C * A_{ss}, \omega_2 = C * A_{sa}, \omega_3 = C * A_{ta}$,where A_{ss}, A_{sa} and A_{ta} are the number of correctly predicted residues on secondary structure (ss), solvent accessibility (sa) and torsion angles (ta), respectively, divided by the total number of the residues on the training proteins. C is the parameter to balance the average magnitude of feature predictions, it was optimized by local experiment samples. In this work, the weighting factors w_1, w_2 and w_3 were set as 3/7, 2/7 and 2/7, respectively.

B. Multiobjective optimization problem

As mentioned above, the sequence search process in this paper can be regarded as a biobjective optimization problem with two analytically unknown fitness functions. In this study, we propose to use a swarm intelligence algorithm to solve this problem.

With two energy functions, the problem can be treated as an optimization process of minimizing two objective functions simultaneously, which can be formulated as:

$$\min_{X\in\Omega} F(X) = (f_1(X), f_2(X))^T$$
$$\begin{cases} X = (x_1, x_2, \cdots, x_n)^T \\ F : \Omega \to \Theta^2 \\ f_i : \Omega \to \Theta \ (\mathrm{i} \ = 1, 2) \end{cases} \tag{11}$$

where X denotes the decision vector, the objective function vector $F(X)$ includes two objective functions, Ω is the feasible domain, Θ^2 denotes the decision space, and $f_i : \Omega \to \Theta$ is the objective function.

In contrast to the single-objective optimization problem, the fitness evaluation is determined by two objective functions; therefore, the fitness value is no longer a one-dimensional value that can be compared in a straightforward manner, and more than one optimal solution may exist. The fitness evaluation for multiobjective optimization problems is usually based on the *Pareto dominance* criterion [28], the detailed definition of which is as Fig. S4.

For multiobjective optimization problems, a decision vector with a minimum value of one objective function may perform relatively poorly in another objective, which makes sense owing to the different characteristics of the objective functions. Different energy functions focus on different aspects of principles of protein folding; thus, protein design methods based on single-objective optimization may have some deficiencies in accurately constructing the model of protein folding. Many heuristic search algorithms can be applied to realize the multi-objective optimization, and in this study, we have designed the particle swarm optimization (PSO) algorithm [28]-based protein design pipeline. The particles representation, initialization, particle position and velocity update, update of the archive and best position, and selection of the final solution can be referred in Algorithm S1 in the supplementary file.

C. Complete process of PSO-driven multi-objective protein design algorithm Hydra

The above sections have introduced the main components of Hydra, including the construction of prior information, sequence space transformation and the biobjective PSO optimization algorithm. To present the detailed and complete flow of our method, the pseudocode of Hydra is provided in the supporting information.

As shown in the pseudocode of Hydra in Algorithm B, it is a *de novo* fixed backbone-based protein design algorithm; therefore, its input should be a protein structure file in PDB format. Before the search process, some necessary preparations should be performed, such as the calculation of the target protein's structural properties, generation of the filtered set (Table S1) for each residue position of the target scaffold, construction and normalization of the structural and statistical information, and sequence space transformation. Then, in the initialization step, all the particles in the PSO are initialized as described in the above sections, the external archive is initialized as the collection of nondominated solutions of all the particles, and the best position is selected according to Eq. (S5). N in line 10 of Algorithm B in the supporting information is the population of particles, and for each particle, $L \times$ *muterate* residue positions will be selected randomly along the target scaffold to update the positions and velocities of these positions. The value of *muterate* will decrease gradually as the iteration increases, which will facilitate the decoy sequence's large-scale variation in the early stage of the search process and maintain the stability and convergence of the decoy sequences in the last stage. This mechanism is critical for the performance of our algorithm. If all residue positions are updated simultaneously in each iteration, the particles will converge prematurely very rapidly, thus not thoroughly exploring the promising area; in contrast, since only a few residue positions are updated in each iteration, as applied in state-of-the-art protein design algorithms, a tremendous number of iterations will be needed to obtain an acceptable solution. Generally, the REMC (Replica Exchange Monte Carlo)-based protein design algorithm normally takes $10^4 \sim 10^5$ orders of magnitude steps to obtain a good solution, while our algorithm only requires $1500 \sim 2000$ iterations to obtain the desired solutions. The position and velocity values of the selected residue positions are then calculated by Eq. (S6) -(S10). After the calculation, the updated position values are projected to the corresponding filtered sets, which maps the value of the transformed sequence space to a real amino acid type. Subsequently, the updated decoy sequences will be evaluated by two objective functions f_1 and f_2. Then, the update of the position and velocity will be accepted or rejected by Eq. (S12). When the inner loop of lines 11–15 ends, the ideal point formulated by Eq. (S13) will be updated. Next, the external archive and best position are updated by the method described in the above section. Every Δ_{KT} iterations, swaps of decoy sequences will take place between adjacent particles that are ranked by their *KT* values. Similarly, every Δ_λ iterations, the sequence space transformation parameters λ_1 and λ_2 are tuned according to Eq. (S18), and then the sequence space is retransformed with the newly tuned parameters. The above iteration continues until the preset maximum of iteration number *It*max is achieved. The final good solutions will be selected from the pareto set.

2.4. Rosetta calculation commands

Protein sequences designed by RosettaDesign in the comparison experiment were obtained from its online server Rosetta.design: http://rosettadesign.med.unc.edu/. The Rosetta 3.10 package can be obtained from https://www.rosettacommons.org/. The Rosetta ab initio has been obtained from https://www.rosettacommons.org/. Ab initio structure predictions were performed by running 'AbinitioRelax.linuxgccrelease' with the commands script listed in Table S10.

To evaluate the Rosetta energy of a protein sequence for a given target structure, we can run 'fixbb.linuxgccrelease' with the protein sequence fixed to obtain a structure with optimized side chain conformations. We then relax the structure by running 'relax.linuxgccrelease'. The final structure is used to calculate the Rosetta energy value.

2.5. Protein expression and purification

The DNA sequences of the designed proteins were synthesized by Shanghai Generay Biotech Company. Expression constructs were created by linking an N-terminal 8 × His-tag, a GST solubility protein tag, and a 3C protease site to the designed sequences in pET-28a vector. Fusion proteins were expressed in *Escherichia coli* BL21 (DE3) by induction with 0.2 mM IPTG at $OD_{600} \sim 0.8$ in M9 minimal medium. Different isotope labels were introduced into the growth medium according to the NMR experimental requirements. Cells were harvested and resuspended in lysis buffer (150 mM NaCl, 50 mM Tris-HCl, pH 8.0) and then lysed by high-pressure homogenizer (Union-Biotech Company). The supernatant was collected by centrifugation at 20 × 000 g and loaded onto the Ni-NTA affinity chromatography column. The fusion protein was eluted with lysis buffer containing 500 mM imidazole. The elution was then cleaved overnight by 3C protease and dialyzed for 4 h with dialysis buffer (150 mM NaCl, 20 mM HEPES, pH 7.5) to remove the imidazole. The target protein was further purified using a second Ni-NTA affinity column to remove the cleaved N-terminal tag followed by gel filtration in a Superdex75 10/300 GL column (GE Healthcare) with GF buffer (50 mM NaCl, 20 mM MES, pH 6.5). Protein fractions were analyzed by SDS-PAGE and concentrated using 3.5 K MWCO concentrators. The final NMR sample included ~ 0.6 mM Hydra-1ubq or Hydra-1r26, 0.5% cocktail protease inhibitor, 0.02% NaN_3, and 10% D_2O. To prepare a sample in 100% D_2O, the regular NMR sample was lyophilized and resolubilized in 100% D_2O. For the CD sample, the size exclusion buffer was replaced with 50 mM KCl, 10 mM K-PO4, pH 7.5.

2.6. Circular dichroism spectroscopy

The designed proteins were first characterized by circular dichroism spectroscopy. A J-715 circular dichroism spectropolarimeter was used for all experiments. Wavelength scans from 190 to 260 nm were conducted. Experimental conditions were 50 mM NaCl, 10 mM K-PO4, pH 7.5 at 298 K. Protein concentrations were approximately 2–4 μM.

2.7. NMR spectroscopy

The NMR experiments were carried out at 25 °C on Agilent DD2 600 MHz or 800 MHz spectrometers equipped with cryogenic probes. The backbone assignments were obtained using the triple resonance experiments, including HNCO, HNCACO, CBCA(CO)NH, HNCACB, HNCA, HN(CO)CA and ^{15}N-edited NOESY-HSQC spectra. These spectra were employed in a nonuniform sampling scheme in the indirect dimensions, and they were reconstructed using the multidimensional decomposition software MDDNMR [29,30] interfaced with NMRPipe [31]. Aliphatic side chain assignments relied on (H)CCH-TOCSY and H(C)CH-TOCSY spectra [32,33]. Aromatic ring resonances were assigned using 3D ^{13}C-edited NOESY spectra. Stereospecific valine and leucine methyl assignments were obtained as described [34] on the basis of the ^{13}C–^{13}C one-bond couplings in a high-resolution 2D ^{1}H–^{13}C HSQC spectrum of 10%-^{13}C, 100%-^{15}N Hydra-1ubq. Complete assignments were obtained for all resonances that appeared in the ^{1}H/^{15}N-HSQC spectrum. A total of 99% of the C and H resonances for all side chains have been assigned.

2.8. Structure calculations

Distance restraints for the structure calculations were derived from cross-peaks in a simultaneous ^{15}N and ^{13}C-NOESY-HSQC (τ_m = 120 ms) [35] and ^{13}C-edited aromatic NOESY-HSQC (τ_m = 120-ms), respectively. The NOE cross-peak assignment was obtained using a combination of manual and automatic procedures. An initial fold of the protein was calculated based on unambiguously assigned NOEs, with subsequent refinement using the NOEassign module implemented in the program CYANA, version 2.1 [36]. Peak analysis of the NOESY spectra was performed by interactive peak picking with the program SPARKY. A total of 136 phi and psi torsion angle restraints were derived from the program TALOS+ [37]. Hydrogen bond restraints were applied only for residues that were clearly in the secondary structure regions as judged by the NOE patterns and chemical shifts and supported by TALOS + . The best 20 of 100 CYANA structures were subjected to molecular dynamics simulation in explicit water using the program CNS [38,39]. The final structures were inspected by PROCHECK [40] and MolProbity [41] using the PSVS software suite [42]. Structures were visualized using the program MOLMOL [43] and Pymol (http://pymol.sourceforge.net, Delano Scientific).

3. Results

The protein targets tested by Hydra in this paper contained 40 proteins downloaded from the RCSB PDB base (https://www.rcsb.org), of which 38 targets were the same as those tested in ABACUS [4]. Forty targets could be divided into 4 groups according to their folding types: all α, all β, α/β and α + β. All the targets were inspected visually in case of irregular shapes and loss of residues. The lengths of the target scaffolds ranged from 75 ~ 191, and their resolutions were better than 1.6 Å. All the target PDB IDs are listed in Table S2.

In this work, we tested designed sequences by computational and wet-lab experiment respectively. For the computational part, we first use Rosetta ab initio structure prediction program to predict the protein structures in a PDB format from the designed and native sequences. Then we compare the predicted structure models using TMscore and R.M.S.D as the metrics. The fast computational tool of Rosetta ab initio structure prediction could provide a rapid initial evaluation of the designed sequences.

3.1. Performance of the biobjective optimization algorithm in Hydra

We first tested the characteristics of our algorithm from the perspective of multiobjective optimization. As a simplification criterion, 4 targets (PDB ID: 2PVB, 1 V05, 1R26, 1UBQ) were selected to be tested as representatives of 4 folding classes, which could show the characteristics of our algorithm comprehensively. The maximum number of iterations of each target was set to 2000, and the population of particles was 8. The remaining detailed settings of the parameters used in Hydra are listed in Table S3.

Fig. 2a illustrates the final Pareto front of 4 targets. The non-dominated solutions of 4 targets were widely distributed, and a considerable amount of the solution was close to the ideal point, which demonstrated that our external archive update mechanism could promote the archive solutions' diverse distribution and convergence to the ideal point. Fig. 2b shows the variation of the best position's objective values during the search process. Generally, the best position of particles varied with the iterations and gradually converged to a stable point. For any target scaffold, both energy function values of the best position were rapidly minimized during the first 300 iterations, owing to the guidance of the offset component in Eq. (S6) (algorithm details in supporting information). Then, the objective values decreased slightly with tiny oscillations due to the attractions of different velocity components, and this process was continued for 900–1200 iterations. At the last stage of the search process, the swarm particles were almost convergent, the distribution of archive solutions tended to be stable; consequently, the best position converged to a notably small area of objective space.

The convergence process of our algorithm is illustrated in Fig. 2c. Without loss of generality, we randomly picked a target (PDB ID: 1UBQ) from 40 targets to show this process. The whole population of particles clearly gradually converged to a stable point despite being diversely initialized. To verify this kind of convergence, all the final sequences of the swarm particles were aligned together, and then we calculated the average entropy of each residue position, which can be formulated as $\sum_{i=1}^{L} p_i \lg p_i$, where L denotes the length of the protein sequence, and p_i is the alignment score at the ith position. If the average entropy value is very low, the sequences of the particles can be regarded as converged. For target 1UBQ, the average entropy value of the final sequences of particles was 0.026, which indicated a very high sequence identity between the final sequences of particles. The convergence of sequences was thus verified in the form of the sequence composition.

3.2. Comparison between multi-objective method and single-objective methods with the same energy functions

To verify the superiority of multiobjective optimization in protein design problem, we designed the sequences of 40 targets by

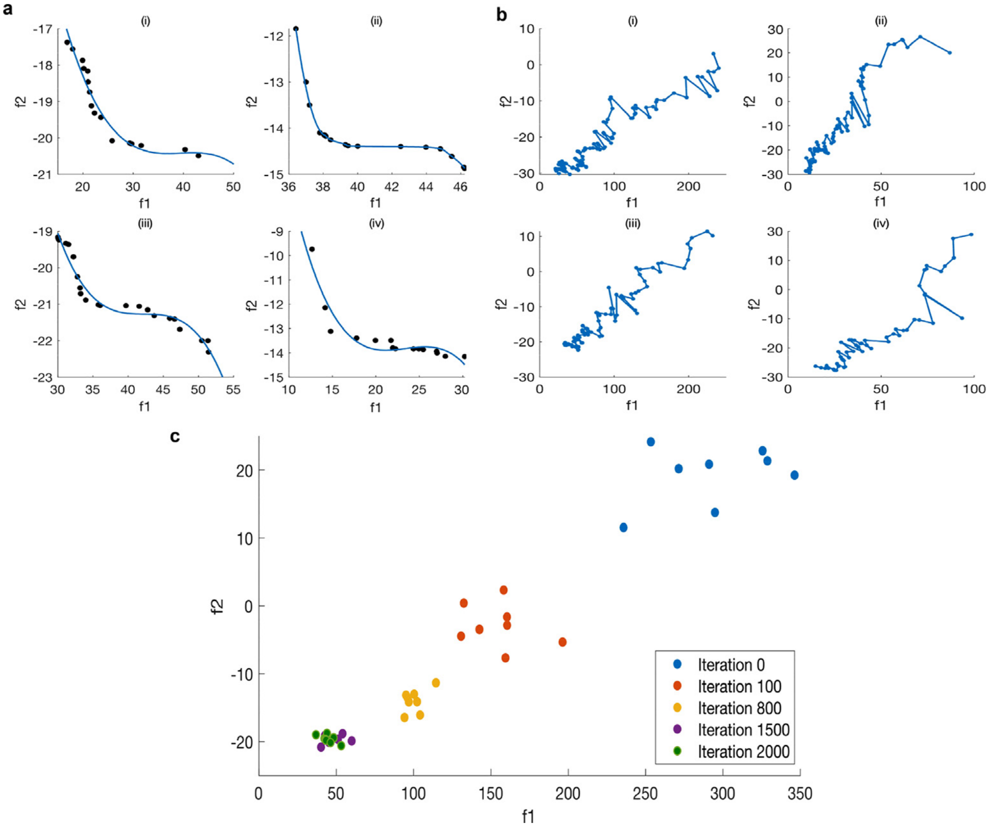


Fig. 2. Performance of the optimization algorithm in Hydra. f_1 is the FoldX energy function, and f_2 is the structure energy function. **a**, Pareto Front of 4 targets in four different folds. (i) Pareto Front of 2PVB. (ii) Pareto Front of 1 V05. (iii) Pareto Front of 1R26. (iv) Pareto Front of 1UBQ. **b**, Variation of the values of the best positions. (i) Variation of the value of the best position of 2PVB. (ii) Variation of the value of the best position for 1 V05. (iii) Variation of the value of the best position for 1R26. (iv) Variation of the value of the best position for1UBQ. **c**, Convergence process of particles in the form of objective values for target 1UBQ.

two methods based on single-objective optimization separately, and these designed sequences were tested by comparison to those designed by Hydra. In this section, we primarily want to show the difference between multi-objective optimization and single-objective optimization, so the energy function should be invariant in the compared methods. Thus, the energy function used in the single-objective methods was as the same as those used in Hydra. The energy function used in the two single-objective methods was the FoldX energy function and structure energy function, respectively. The single-objective optimization algorithm's architecture was almost the same as Hydra, except that the decoy sequences were evaluated by a single energy function, and the external archive mechanism was therefore abandoned. All the parameters of the algorithm and experiment were identical to those of Hydra.

Fig. 3a shows that the results of Hydra were slightly better than those of two single-objective based methods in terms of secondary structure assignment and sequence identity, which fulfils our expectations considering the cooperation of three kinds of information in Hydra. To further test the foldability of sequences designed by single-objective methods, we also applied the Rosetta *ab initio* structure prediction to obtain the corresponding structures of the designed sequences for the same 40 targets. As illustrated in Fig. 3b, compared with the Single FoldX and Single Structure, the sequences designed by Hydra achieved the best performance in terms of the fraction of highly target-like models and average highest TM-score. Especially in all α and all β fold classes, the correct foldability of sequences designed by Hydra significantly surpassed that of sequences designed using single-objective based methods significantly. From Fig. 3b, it is clear that Hydra possessed the best performance in all fold classes, which indicated that the multiobjective mechanism could enhance the correct foldability of the designed sequences.

We also have compared Hydra with the single energy function method that composed by weighted sum of FoldX and structure energy terms. The weighted energy function is formulated as $E = E_{FoldX} + \eta E_{Structure}$. The weight η is adjusted to make the two energy terms have comparable contributions, in this work it is set to 1.37 according to local tests. All the other parameters are kept the same for the two methods. From the results, Hydra has achieved better performance than the standard weighted single energy function. The results can be referred in Fig. S5 in Supplementary file.

Since Hydra outperformed single-objective methods which applied FoldX, Structure energy function, and the weighted sum of the two energy functions respectively. And the mechanism of the three single-objective methods was almost as same as that of Hydra, suggesting that the multi-objective method has brought the improvement. The reason for this improvement may be that optimizing two energy functions simultaneously can help reducing the local minimal solutions, and two energy functions can make the result more robust.

3.3. Tests of protein sequences designed by Hydra and other methods

We have tested the sequences of 40 targets designed by our method Hydra. All the parameter settings were the same as those mentioned in the above section (2000 iterations, 8 particles), and the details can be found in Table S3. For comparison, we also collected the sequences designed by the ABACUS [4] and Rosetta fixed backbone design [3], of which the sequences designed by ABACUS

a

Comparison with single-objective method

Method	**SS error**	**Sequence identity**
Single FoldX	0.25	0.2913
Single Structure	0.22	0.2877
Hydra	0.20	0.3080

Single FoldX: single-objective method with FoldX energy function.
Single Structure: single-objective method with structure energy function.

b

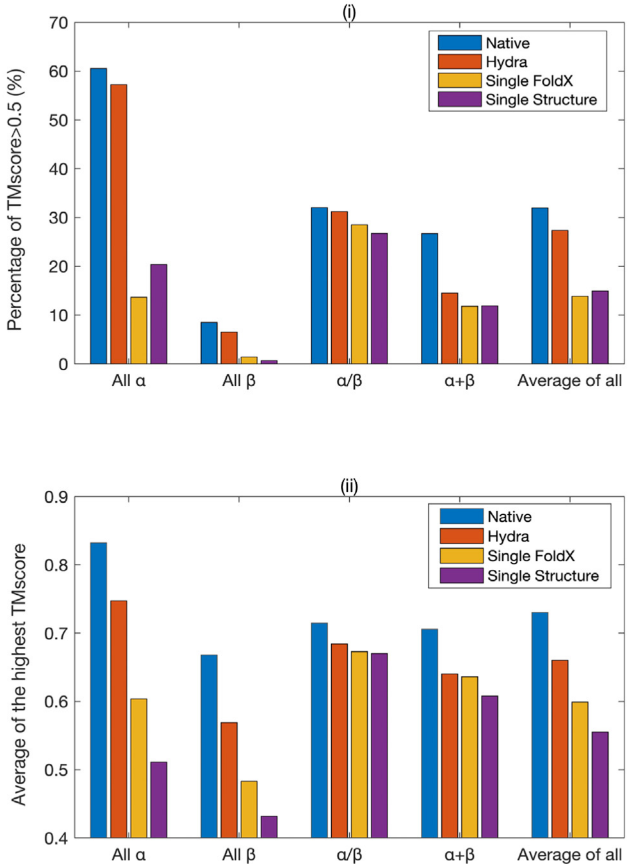
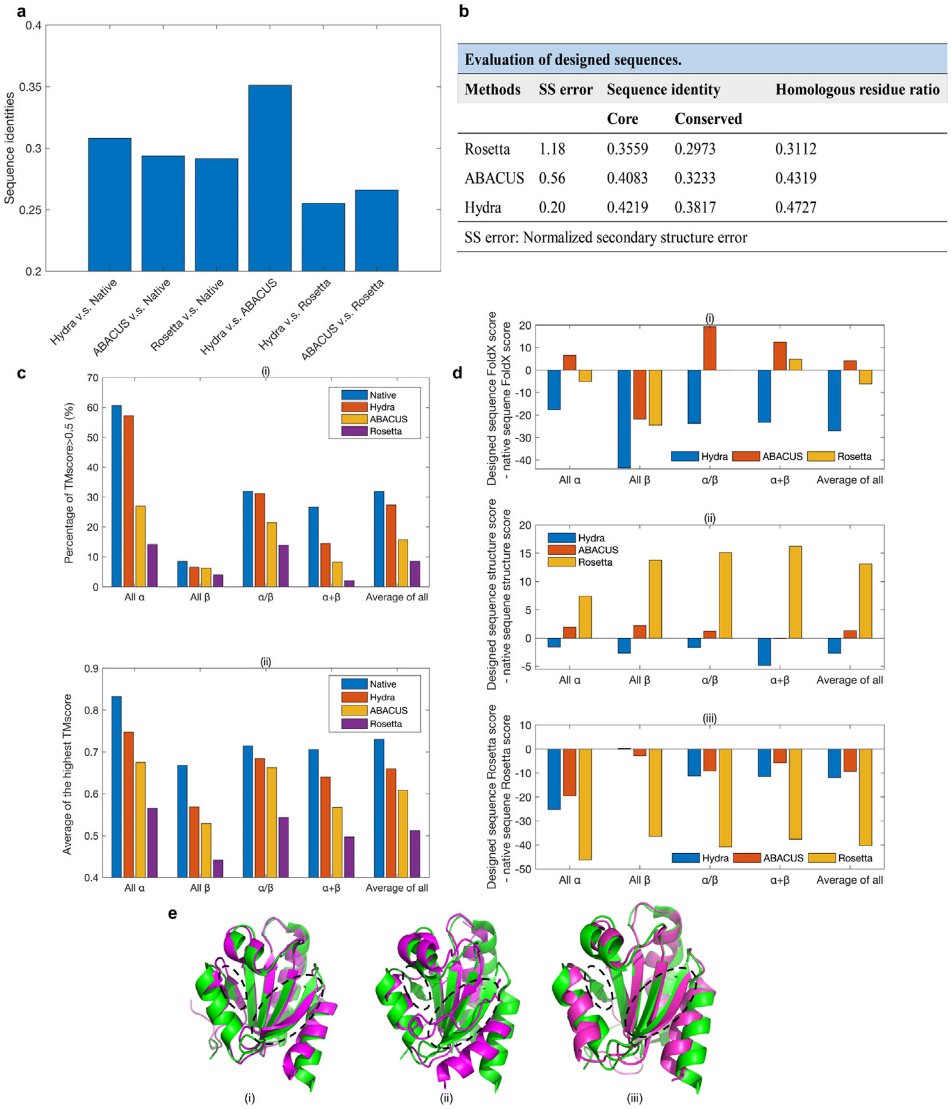


Fig. 3. Comparison with single-objective methods. **a**, Performance in terms of secondary structure and sequence identity. **b**, Results of the structure models *ab initio* predicted from sequences designed by single-objective methods. (i) Fraction of highly target-like (*TM-score > 0.5*) models in structures predicted from native sequences and sequences designed by single-objective methods. (ii) Average highest TM-score values of structures predicted from native sequences and sequences designed by single-objective methods.

were directly selected from the supplementary file of its paper, except the sequences of two newly added targets that were obtained by ABACUS standalone software, and the sequences designed by Rosetta were obtained from its online server http://rosettadesign.med.unc.edu/login.php. All the designed protein sequences of 40 targets using the three methods are listed in Table S4–S6.

Sequence identity values of the average of 40 targets are compared in Fig. 4a. Overall, the sequence identity of our method relative to the respective native sequence was 0.3080, which was slightly higher than ABACUS's 0.2936 and Rosetta's 0.2916. Interestingly, the sequence identity between sequences designed by Hydra and Rosetta was notably low; therefore, our method and Rosetta could complement each other. The sequence identities of core (solvent less exposed) positions and conserved (residue positions with low entropy in the multiple sequence alignment based on PSI_BLAST) positions of Hydra respectively were 0.4219 and 0.3817, both of which were higher than ABACUS's 0.4083 and 0.3233 and Rosetta's 0.3559 and 0.3233 (Fig. 4b). A high sequence identity in core and conserved positions is of great significance because these positions contribute more to folding. We also calculated the ratio of highly homologous residues, which are the residues of the designed sequence with a Blosum62 mutation score higher than 1 relative to the respective residues of the native sequence in this study, and the ratio of Hydra, ABACUS and Rosetta respectively was 0.4727, 0.4319 and 0.3112 (Fig. 4b).

To go a step further, we also use CLUSTALW to do sequence alignment to explore the similarity between native sequences and designed sequences. The result shows that the average alignment score between native sequences and designed sequences is 39.4745 for Hydra, 28.9474 for ABACUS and 32.8947 for Rosetta. It seems that on our tested proteins, sequences designed by Hydra prefer to have more similarities with native sequences. According to the designed 1UBQ sequences, the maximum sequence identity observed to native sequence is 0.4079. For other experimented proteins, the maximum sequence identity observed to native sequence is 0.55 for 1R26. Our experimental results indicate interestingly that even different sequences with no >0.5 sequence iden-



Evaluation of designed sequences.

Methods	SS error	Sequence identity		Homologous residue ratio
		Core	Conserved	
Rosetta	1.18	0.3559	0.2973	0.3112
ABACUS	0.56	0.4083	0.3233	0.4319
Hydra	0.20	0.4219	0.3817	0.4727

SS error: Normalized secondary structure error

Fig. 4. Computational performance of Hydra, ABACUS and Rosetta. **a**, Sequence identities between native sequences, protein sequences designed by Hydra, ABACUS, and Rosetta fixed backbone design. **b**, Evaluation of designed sequences. **c**, Results of the structure models *ab initio* predicted from native sequences and designed sequences. (i) Fraction of highly target-like (TM-score > 0.5) models in structures predicted from native sequences and sequences designed by different methods. (ii) Average highest TM-score values of structures predicted from native sequences and sequences designed by different methods. **d**, Average energies of designed sequences relative to corresponding native sequences for 40 targets. (i) Average FoldX energies of designed sequences relative to corresponding native sequences. (ii) Average structure energies of designed sequences relative to corresponding native sequences. (iii) Average Rosetta energies of designed sequences relative to corresponding native sequences. **e**, Illustration of structures predicted from different designed sequences by three methods for target 1R26. (i) Native structure of 1R26 (green) and best predicted structure model from the sequence designed by Hydra (purple). (ii) Native structure of 1R26 (green) and best predicted structure model from the sequence designed by Rosetta (purple). (iii) Native structure of 1R26 (green) and best predicted structure model from the sequence designed by ABACUS (purple). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

tity could fold to very similar structure. Combining these results, we could find that the similarity of designed sequence to some extent could provide a clue for the following folding. We have also provided the amino acid composition of native sequences and sequences designed by Hydra, ABACUS and Rosetta as a reference (Table S8).

To further test our method, we examined the accuracy of the secondary structure of the designed sequences and compared them with the native sequences. The secondary structures of the designed sequences were predicted by PSIPRED [44], and those of the native sequences were assigned by Stride [24]. To account for the inherent error induced by the inaccuracy of PSIPRED, we applied the previously proposed definition of the normalized secondary structure error (NSSE) [13], which is formulated as $NSSE = (EDS - ETS)/ETS$, where *EDS* is the secondary structure error of the designed sequence relative to the Stride file of the native sequence, and *ETS* is that of the native sequence. NSSE of designed sequence by Hydra was 0.20, whereas that of ABACUS and Rosetta was considerably higher at 0.56 and 1.18, respectively (Fig. 4b). We have used the Spider3 to compute the solvent accessibility (SA) of 2PVB sequence designed by Hydra (refer to Table S4) and native sequence respectively. The results show that Hydra sequence has ~ 9% more core residues (SA < 25) than native sequences. And we also have used I-TASSER [13] to compute the structure model of 2PVB sequence designed by Hydra. Then we have used STRIDE to calculate the SA of the native and computed structure models, the results show that computed Hydra structure has ~ 2% more core residues than the target structure, which indicates that Hydra structure is slightly better packed than the native structure from the computational point of view. Clearly, full use of different kinds of information in our method could promote the interaction between the secondary structure propensity and overall foldability of the designed sequences. In our local experiments, we have observed that accurate secondary structure is the basis of accurate tertiary structure. As it is shown in the next part, sequences designed by Hydra have good secondary structure accuracy, which makes a good foundation for packing. These results potentially indicate that Hydra can successfully make full use of the overall and local information of the target structure.

We applied Rosetta *ab initio* structure prediction to test the foldability of the sequences designed using different methods, and as a reference, the native sequence of each target was also used to predict its structures by Rosetta *ab initio* structure prediction. In the Rosetta *ab initio* structure prediction, none of the homologous native protein structures was used as template, which ensured the rationality of this test method. For each sequence, 200 tertiary structures were predicted. Usually, higher accuracy will be achieved with more models generated. From our local tests, 200 models could result in relatively reliable result just as similar observation in the work of Xiong, P.et al [4]. Considering the balance between the performance and computation time, we generated 200 models in this work and will test more in the future on more powerful computer system. The similarity of the predicted structures with corresponding target scaffolds was evaluated by the TM-score. The structures could be regarded as the same folds with their respective targets if the TM-score was higher than 0.5. For each target, we calculated the fraction of highly target-like (TM-score > 0.5) models among 200 predicted ones, and the highest TM-score value among 200 predicted structures was also counted. All the results for 40 targets are listed in Table S7. The average results of the targets of different fold classes are shown in Fig. 4c.

As expected, we found that the native sequences achieved the best performance in both highly target-like fraction and average highest TM-score, which verified the rationality of the test method based on Rosetta *ab initio* prediction. From Fig. 4c(i), the sequences designed by Hydra generally led to high fractional target-like models as native sequences, significantly surpassing the sequences designed by ABACUS and Rosetta. Especially for all α and α/β fold classes, the score of Hydra was almost as high as that of native sequences, indicating a remarkable foldability in these fold classes. Although the sequences designed using Hydra did not achieve the same high fraction of target-like models as the native sequences in the $\alpha + \beta$ fold class, our method still outperformed ABACUS and Rosetta. For all β targets, the sequences designed by the three methods led to a highly target-like fraction lower than 10%, as did the native sequences. This result may have been observed because Rosetta *ab initio* structure prediction cannot predict the structures of all β sequences accurately or because accurate structure prediction of all β sequences is rather difficult. Fig. 4c(ii) shows the average highest TM-score of sequences designed by different methods and native sequences. Overall, the results using our method were slightly better than those achieved with ABACUS and surpassed those of Rosetta. For each fold class, Hydra possessed the highest average TM-score compared with the other two methods, especially for the α/β class, and Hydra achieved a comparable value to that of the native sequences. The fraction of highly target-like models denoted the probability that the sequence could be folded into the target scaffold, and the average highest TM-score relative to the target scaffold indicated the degree of similarity between the folded structure and target scaffold, both of which were important to evaluate the foldability of the designed sequences. Our method surpassed ABACUS and Rosetta in both evaluation metrics. A detailed analysis of the results of the designed sequences is listed in Table S7.

To examine the complementarity and interactions between different kinds of energy functions, we employed the FoldX energy function, structure energy function and Rosetta energy function to evaluate the sequences designed by Hydra, ABACUS and Rosetta separately. The corresponding native sequences were also calculated with the same energy functions. As expected, the sequences designed by Hydra had low FoldX values in all the fold classes, and the values of the sequences designed by Rosetta overall were also lower than those of corresponding native sequences, which was due to the physics of the force field included in the Rosetta energy function similar to FoldX (Fig. 4d(i)). The structure energy values of the sequences designed by Hydra were lower than those of the corresponding native sequences, while for sequences designed by ABACUS and Rosetta, the values were higher. Especially for the sequences designed by Rosetta, the structure energy values were notably higher compared with those of the corresponding native sequences, primarily because no information about the target's homologous structures was included in the Rosetta energy function. Considering that the native sequences could be absolutely folded into the target scaffolds, since the structure energy function favored native sequences over the sequences designed by Rosetta, the structure energy function captured some criteria of foldability that were not included in the Rosetta energy function. Intriguingly, Fig. 4d(iii) shows that the Rosetta energies of the sequences designed by Hydra were lower than those of the native sequences, validating the robustness and reliability of the sequences designed by Hydra on the one hand and indicating that homologous structure information can complement the Rosetta energy function in the protein design on the other hand. From Fig. 4d(i)-(iii), we found that all three energies of the sequences designed by Hydra were lower than those of the corresponding native sequences, revealing good performance and efficiency of our biobjective optimization algorithm for sufficiently minimizing the two energy functions with relatively few iterations. We have also generated Energies vs RMSD over C-alpha atoms plot to show the energy landscapes of Hydra. The plot and analysis are shown in Fig. S6 in supplementary file.

For some targets, the average sequence identity between sequences designed by Hydra and target sequences was very similar to that of ABACUS and Rosetta, and our Hydra method still outperformed ABACUS and Rosetta. An example showing the different foldabilities of the sequences designed by three methods is shown (Fig. 4e). The 1R26 is the thioredoxin from *Trypanosoma brucei brucei*, and it is an α/β fold class protein with three alpha helices and five beta strands, where one of the beta strands is very short and connected by two segments of coil structure. The TM-score of the predicted structures in Fig. 4e(i)–(iii) relative to 1R26′s native structure was 0.8511, 0.6441 and 0.8093, respectively. Comparison of the structures in Fig. 4e(i) and e(ii) showed that the structure predicted from the sequence designed by Hydra fit the target structure much better than that of Rosetta. Particularly, the two segments of the beta strand marked by black dotted lines were mistakenly folded into coils in Fig. 4e(ii), while they were accurately folded in the structure predicted from the sequence designed by Hydra in Fig. 4e(i). Compared with Fig. 4e(iii), it can be observed that the structure of Hydra fit the target scaffold better than that of ABACUS, especially for the three alpha helices, potentially due to the homologous structure information of the target included in Hydra, which could capture the correlation between different local structures to facilitate the packing of proteins.

Interestingly, the short beta strand marked by black dotted lines was the only correctly folded structure in the sequence designed by Hydra among the three methods. The difficulty of recognizing the complex topology information related to the short beta strand connected by two coil segments supported the accuracy and robustness of the Hydra Design method. Through Rosetta *ab initio* structure prediction verification, the foldabilities of the sequences designed by Hydra were demonstrated, since the Rosetta energy function was not applied in our method, thereby ensuring the reliability of the experimental results of Hydra.

3.4. Structural verification of the designed protein

To validate the designed proteins, we experimentally characterized two randomly selected proteins, thioredoxin (PDB ID: 1R26) and ubiquitin (PDB ID: 1UBQ) (Fig. S1) [45,46]. For the designed sequences Hydra-1r26 and Hydra-1ubq, constructs containing an N-terminal 8 × His-tag and a GST tag were synthesized and cloned into the expression vector pET-28a, which was then transformed into *Escherichia coli* BL21 (DE3). The expressed proteins were further purified by Ni-NTA affinity and size exclusion chromatography and characterized by circular dichroism (CD) and solution NMR spectroscopy. As shown in Fig. 5 and Fig. S2, the designed proteins were successfully expressed and purified, generating good CD spectra and ^{1}H-^{15}N heteronuclear single quantum coherence (HSQC) spectra, indicating that the proteins were well folded. We also have found that there does exist a *cis*-Pro (Table S4 in Supplementary File) in our designed sequence for 1R26. Since structure is the only input in Hydra, there is no native sequence composition information used during the whole process, and the *cis*-Pro is not specifically kept in the design of 1R26 using Hydra. Thus, the remained feature in sequence indicates that Hydra relatively could maintain the key functionality of protein. We chose Hydra-1ubq, which showed better NMR spectrum quality (Fig. S2), to further solve the NMR structure as an example of the protein design.

The structure determination was carried out following standard NMR protocols. Briefly, sequence specific backbone assignment was accomplished using the triple resonance experiments (Fig. 5a and Fig. S3), and the assigned chemical shift values were used as input for the TALOS + program to obtain the secondary structures [37]. Then, distance restraints derived from nuclear Overhauser enhancements (NOEs) were used to determine the structures (details in Methods). A total of 100 structures were calculated, and the 20 structures with the lowest energy were selected as the final structural ensemble (Fig. 5b). The 20 structures converged very well, with a root mean squared deviation (backbone RMSD) of 0.3 Å and 0.7 Å for the backbone and all heavy atoms, respectively. The structure alignment of the NMR structure of Hydra-1ubq (PDB ID: 6L0L, BMRB ID: 36289) and the crystal structure (PDB ID: 1UBQ) showed a 1.074 Å backbone RMSD (Fig. 5c). Both structures possessed five beta-strands and one helix packed in a sandwich fold (Fig. 5c). The high structural similarity between the designed protein Hydra-1ubq and the original protein ubiquitin demonstrated that the new protein sequence designed from Hydra achieved the desired structure. The information used for structural calculation and the detailed structure statistics are listed in Tables S11 and S12.

4. Conclusion and discussions

In this paper, we proposed a new *de novo* protein design method named Hydra. This method is based on multi-objective PSO swarm intelligence algorithm, and the two energy functions

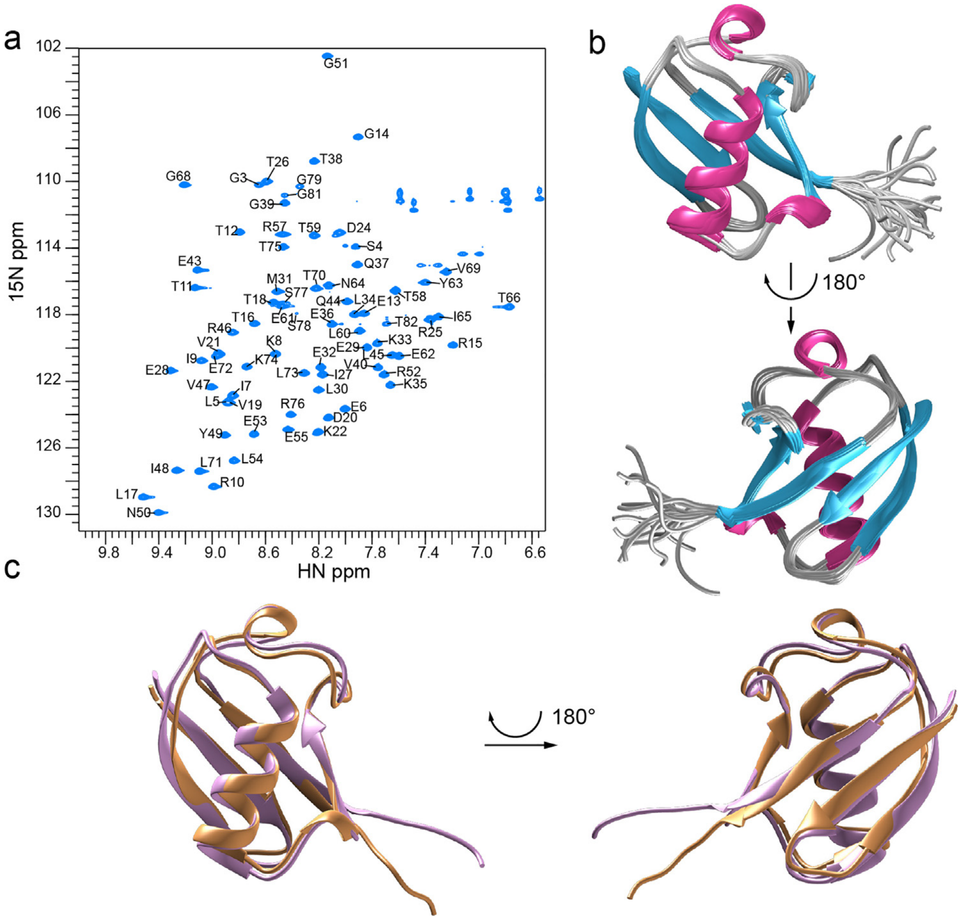


Fig. 5. Structural characterization of the designed protein Hydra-1ubq. **a**. The ^{1}H-^{15}N HSQC spectrum of the ^{15}N-^{13}C labeled Hydra-1ubq sample recorded on an Agilent spectrometer operating at a ^{1}H frequency of 600 MHz. **b**. Thin ribbon representation of the ensemble of 20 lowest energy structures of Hydra-1ubq. **c**. Three-dimensional structure superposition of the NMR structure of Hydra-1ubq (plum) and the crystal structure ubiquitin (PDB ID: 1UBQ, sandy brown).

are the FoldX force field and a structure energy function containing some local structural properties. Moreover, the statistical information and information extracted from the target's homologous structure family were also applied to conduct the sequence space transformation, an innovative method proposed in this paper to reduce the search scope. Convergence analysis of our algorithm is also presented and verified by our experiment results. This newly developed protein design method demonstrates that a multiobjective optimization method could improve the foldability of the designed sequences compared with single-objective based methods and that the latent quantitative information between different amino acid types will improve the performance of the algorithm.

There are two features of the developed software as follows:

(1) In Hydra, two energy functions in our algorithm are minimized simultaneously in a multiobjective manner during the search process. The two energy functions belong to different categories based on the physics of the force field and structural properties, respectively. In contrast to methods based on single-objective optimization, Hydra combines the characteristics of two different energy functions and thus can capture the mapping relationship between protein structure and sequence more accurately. Conversely, Hydra is more robust owing to the combination of multiple energy functions that can promote and restrain each other to prevent particles from falling into local minima during the search process.
(2) We introduced a sequence space transformation procedure in our algorithm to simplify the search process. The space transformation procedure has two advantages: the scope of the search space for good solutions is largely reduced, and the original discrete optimization problem can be solved as a continuous optimization problem in the transformed sequence space.

In our design of the Hydra pipeline, we also systematically tested the importance of the heuristic search step. If the optimization search is optional, then when we randomly select the amino acid type with high fusion value in Eq. (3) (Methods) for each residue position, the generated sequence can be a good solution for the target scaffold with high possibilities and without an extensive search process. To test this possibility, we generated many sequences using this random selection method. However, the energy function values of almost all the generated sequences were considerably higher than those of the sequences designed by heuristic Hydra, and very few of the generated sequences were correctly folded into the corresponding target scaffolds with a TM-score > 0.5 because a large amount of space must still be explored in the transformed sequence space and the randomly generated sequences do not account for the physical constraints. Clearly, the search process is critical to our method because it guides the particles in a smoothing way with the full cooperation of different kinds of information, and it can predict the direction, possibly leading to better solutions in the continuous transformed space for subsequent moves. Thus, our algorithm can prevent particles from falling into local minima in a relative manner.

Compared with multiple sequence alignment-guided consensus design methods, Hydra is a fixed backbone protein design method, where the input is the target peptide, and the output is the designed protein sequence. We do not consider the input peptide is whether stable or not, at current stage, what this program concerns about is to get the sequence that can fold to an expected structure. To the best of our knowledge, most of the multiple sequence alignment-guided consensus design methods solve this problem from a relatively different point of view, whose input can include the target peptide and the corresponding sequence. It exploits amino acid conservation in sets of homologous proteins to identify likely beneficial mutations to stabilize the original protein. Actually, they solve relatively different problems, but these two pipelines can be complementary. Hydra can produce reliable sequence that could fold to a certain structure, while consensus design could explore more stable peptides. We could combine these two kinds of methods to perform them alternately, the much more stable and accurate sequences could be produced. For exampple, we could use Hydra to output a foldable sequence for a peptide, then the consensus design could do some beneficial mutations in this sequence to make the original peptide more stable. After a certain number of iterations, we could get the sequence that well folded to a more stable peptide compared to the original one.

In future work, we will further dig out the potential relationship between the number of particles in the algorithm and the convergence stability of the searched sequence as our local tests have shown the performance will be related to this parameter, which would have impact on the search diversity (Fig. S7). We will also try to apply more energy functions in the algorithm to make the method more robust and accurate. Furtherly, we could also make use of the recent protocols based on deep learning to coordinate with Hydra, which makes a remarkable progress in protein design. In the help of super computers, much more particles could be run to produce better results. We will also try to do more experimental tests on both designed and native sequence in order to explore the differences in terms of expression, solubility and thermostability, especially to see if the key sites of the protein are retained. Not only the structure, but functionality will also be considered in experiment.

Author contributions

H. S. and R. L. conceived the study. B. O. conceived the NMR experiments. N. Z. prepared NMR samples, collected and analyzed NMR data with the help of B. O. and B. W.. H.S., R. L. and B. O. wrote the manuscript with help from all authors.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (No. 2018YFC0910500), National Natural Science Foundation of China (No. 61725302, 62073219), and Science and Technology Commission of Shanghai Municipality (No.17JC1403500) to H. S.. This work was also supported by grants from a National key R&D Program of China (2017YFA0504804), Key Research Program of Frontier Sciences, CAS (QYZDB-SSW-SMC043) and the Chinese Academy of Sciences (CAS) Major Science and Technology Infrastructure Open Research Projects to B. O.. We also thank the staffs from Nuclear Magnetic Resonance System/ Mass Spectrometry System at National Facility for Protein Science in Shanghai, Zhangjiang Laboratory (NFPS, ZJLab), China for providing technical support and assistance in data collection and analysis.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2021.04.046.

References:

[1] Gräslund S, Nordlund P, Weigelt J, Hallberg BM, Bray J, Gileadi O, et al. Protein production and purification. Nat Methods 2008;5(2):135.

[2] Chevalier A, Silva D-A, Rocklin GJ, Hicks DR, Vergara R, Murapa P, et al. Massively parallel de novo protein design for targeted therapeutics. Nature 2017;550(7674):74–9.

[3] Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. Design of a novel globular protein fold with atomic-level accuracy. Science 2003;302 (5649):1364–8.

[4] Xiong P, Wang M, Zhou X, Zhang T, Zhang J, Chen Q, et al. Protein design with a comprehensive statistical energy function and boosted by experimental selection for foldability. Nat Commun 2014;5(1). https://doi.org/10.1038/ncomms6330.

[5] Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 1983;4(2):187–217.

[6] Rohl CA, Strauss CE, Misura KM, Baker D. Protein structure prediction using Rosetta. Methods Enzymol 2003;383(383):66.

[7] Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci. 2002;11(11):2714–26.

[8] Oldziej S, Czaplewski C, Liwo A, Chinchio M, Nanias M, Vila JA, et al. Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field: assessment in two blind tests. Proc Natl Acad Sci 2005;102 (21):7547–52.

[9] Case DA, Cheatham TE, Darden T, Gohlke H, Luo R, Merz KM, et al. The Amber biomolecular simulation programs. J Comput Chem. 2005;26 (16):1668–88.

[10] Belure SV, Shir OM, Nanda V. Protein design by multiobjective optimization: evolutionary and non-evolutionary approaches. In: Proceedings of the genetic and evolutionary computation conference. p. 1081–8.

[11] Zeugmann T, Poupart P, Kennedy J, Jin X, Han J, Saitta L, et al. In: Encyclopedia of machine learning. Boston, MA: Springer US; 2010. p. 760–6. https://doi.org/10.1007/978-0-387-30164-8_630.

[12] Schymkowitz JWH, Rousseau F, Martins IC, Ferkinghoff-Borg J, Stricher F, Serrano L. Prediction of water and metal binding sites and their affinities by using the Fold-X force field. Proc Natl Acad Sci 2005;102(29):10147–52.

[13] Mitra P, Shultis D, Brender JR, Czajka J, Marsh D, Gray F, et al. An evolution-based approach to de novo protein design and case study on Mycobacterium tuberculosis. PLoS Comput Biol 2013;9(10):e1003298.

[14] Brender JR, Shultis D, Khattak NA, et al. An evolution-based approach to de novo protein design. New York, NY: Computational Protein Design. Humana Press; 2017. p. 243–64.

[15] Zhang Y, Skolnick JJ. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33(7):2302–9.

[16] Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins Struct Funct Bioinf 2004;57(4):702–10.

[17] Xu J, Zhang YJ. How significant is a protein structure similarity with TM-score= 0.5?. Bioinformatics 2010;26(7):889–95.

[18] Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D, Goodsell DS, et al. The RCSB Protein Data Bank: redesigned web site and web services. Nucleic Acids Res 2011;39(Database):D392–401.

[19] Edgar RC, Batzoglou S. Multiple sequence alignment. Curr Opin Struct Biol 2006;16(3):368–73.

[20] Gribskov M, McLachlan AD, Eisenberg D. Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci 1987;84(13):4355–8.

[21] Eddy SR. Where did the BLOSUM62 alignment score matrix come from?. Nat Biotechnol 2004;22(8):1035–6.

[22] 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 Res 1997;25(17):3389–402.

[23] Saravanan KM, Balasubramanian H, Nallusamy S, Samuel S. Sequence and structural analysis of two designed proteins with 88% identity adopting different folds. Protein Eng Des Sel 2010;23(12):911–8.

[24] Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins Struct Funct Bioinf 1995;23(4):566–79.

[25] Coluzza I. Computational protein design: a review. J Phys: Condens Matter 2017;29(14):143001. https://doi.org/10.1088/1361-648X/aa5c76.

[26] Krivov GG, Shapovalov MV, Dunbrack RL. Improved prediction of protein side-chain conformations with SCWRL4. Proteins Struct Funct Bioinf 2009;77 (4):778–95.

[27] Heffernan R, Paliwal K, Lyons J, Singh J, Yang Y, Zhou Y. Single-sequence-based prediction of protein secondary structures and solvent accessibility by deep whole-sequence learning. J Comput Chem 2018;39(26):2210–6.

[28] Kennedy J, Eberhart R. Particle swarm optimization. Proceedings of ICNN'95-international conference on neural networks. IEEE, 1995;4:1942–1948.

[29] Kazimierczuk K, Orekhov VY. Accelerated NMR spectroscopy by using compressed sensing. Angew Chem Int Ed Engl 2011;50(24):5556–9.

[30] Orekhov VY, Jaravine VA. Analysis of non-uniformly sampled spectra with multi-dimensional decomposition. Prog Nucl Magn Reson Spectrosc 2011;59 (3):271–92.

[31] Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. J Biomol NMR 1995;6(3):277–93.

[32] Bax A, Vuister GW, Grzesiek S, Delaglio F, Wang AC, Tschudin R, et al. Measurement of homo- and heteronuclear J couplings from quantitative J correlation. Methods Enzymol 1994;239:79–105.

[33] Kay LE. NMR methods for the study of protein structure and dynamics. Biochem Cell Biol 1997;75(1):1–15.

[34] Neri D, Szyperski T, Otting G, Senn H, Wuthrich K. Stereospecific nuclear magnetic resonance assignments of the methyl groups of valine and leucine in the DNA-binding domain of the 434 repressor by biosynthetically directed fractional 13C labeling. Biochemistry 1989;28(19):7510–6.

[35] Pascal SM, Muhandiram DR, Yamazaki T, et al. Simultaneous acquisition of 15N-and 13C-edited NOE spectra of proteins dissolved in H2O. J Magn Resonance Ser B (Print) 1994;103(2):197–201.

[36] Guntert P. Automated NMR structure calculation with CYANA. Methods Mol Biol 2004;278:353–78.

[37] Shen Y, Delaglio F, Cornilescu G, Bax Ad. TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J Biomol NMR 2009;44(4):213–23.

[38] Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, et al. Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr. 1998;54(Pt 5):905–21.

[39] Linge JP, Williams MA, Spronk CAEM, Bonvin AMJJ, Nilges M. Refinement of protein structures in explicit solvent. Proteins 2003;50(3):496–506.

[40] Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R, Thornton JM. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J Biomol NMR 1996;8(4):477–86.

[41] Lovell SC, Davis IW, Arendall 3rd WB, de Bakker PI, Word JM, Prisant MG, et al. Structure validation by Calpha geometry: phi, psi and Cbeta deviation. Proteins 2003;50(3):437–50.

[42] Bhattacharya A, Tejero R, Montelione GT. Evaluating protein structures determined by structural genomics consortia. Proteins 2007;66(4):778–95.

[43] Koradi R, Billeter M, Wuthrich K. MOLMOL: a program for display and analysis of macromolecular structures. J Mol Graph. 1996;14(1):51–55, 29–32.

[44] McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinformatics 2000;16(4):404–5.

[45] Friemann R, Schmidt H, Ramaswamy S, Forstner M, Krauth-Siegel RL, Eklund H. Structure of thioredoxin from Trypanosoma brucei brucei. FEBS Lett 2003;554(3):301–5.

[46] Vijay-Kumar S, Bugg CE, Cook WJ. Structure of ubiquitin refined at 1.8 A resolution. J Mol Biol 1987;194(3):531–44.