Received: 2 November 2021 | Revised: 8 November 2022 | Accepted: 6 February 2023

DOI: 10.1002/prot.26478

Protein structure prediction using the evolutionary algorithm USPEX

Pavel Rachitskii[1] | Ivan Kruglov[1,2] | Alexei V. Finkelstein[3,4,5] | Artem R. Oganov[6]

[1]Moscow Institute of Physics and Technology, Dolgoprudny, Russia

[2]Dukhov Research Institute of Automatics (VNIIA), Moscow, Russia

[3]Institute of Protein Research of the Russian Academy of Sciences, Moscow, Russia

[4]Biology Department of the Lomonosov Moscow State University, Moscow, Russia

[5]Biotechnology Department of the Lomonosov Moscow State University, Moscow, Russia

[6]Skolkovo Institute of Science and Technology, Skolkovo Innovation Center, Moscow, Russia

Correspondence
Pavel Rachitskii, Moscow Institute of Physics and Technology, Dolgoprudny 141700, Russia.
Email: rachitskiy.pyu@phystech.edu

Abstract

Protein structure prediction is one of major problems of modern biophysics: current attempts to predict the tertiary protein structure from amino acid sequence are successful mostly when the use of big data and machine learning allows one to reduce the "prediction problem" to the "problem of recognition". Compared with recent successes of deep learning, classical predictive methods lag behind in their accuracy for the prediction of stable conformations. Therefore, in this work we extended the evolutionary algorithm USPEX to predict protein structure based on global optimization starting with the amino acid sequence. Moreover, we compared frequently used force fields for the task of protein structure prediction. Protein structure relaxation and energy calculations were performed using Tinker (with several different force fields) and Rosetta (with REF2015 force field) codes. To create new protein structure models in the USPEX algorithm, we developed novel variation operators. The test of the new method on seven proteins having (for simplicity) no cis-proline (with $\omega \approx 0°$) residues, and a length of up to 100 residues, revealed that our algorithm predicts tertiary structures of proteins with high accuracy. The comparison of the final potential energies of the predicted protein structures obtained using the USPEX and the Rosetta Abinitio approach showed that in most cases the developed algorithm found structures with close or even lower energy (Amber/Charmm/Oplsaal) and scoring function (REF2015). While USPEX has clearly demonstrated its ability to find very deep energy minima, our study showed that the existing force fields are not sufficiently accurate for accurate blind prediction of protein structures without further experimental verification.

KEYWORDS

evolutionary algorithm, protein folding, protein structure prediction, USPEX, variation operator

1 | INTRODUCTION

1.1 | Motivation

Proteins—organic compounds consisting of amino acid residues—perform many functions in living organisms. For example, collagen is very important for skin, ligaments, and tendons, providing strength and elasticity. In bones and teeth, it is mineralized to form hard tissues, contributing to their bearing capacity. Some proteins, such as myosin or kinesin, can be considered as molecular machines that can move in a "quasi-mechanical" way. These multiprotein complexes frequently perform vital functions, such as muscle contraction and gene expression. In addition, proteins are responsible for many other processes, from catalysis of biochemical reactions to signal transmission between cells. The function of a protein is defined by its structure, and knowledge of this structure helps understand the function and mechanism of the protein's operation, providing the basis for the creation of new effective drugs.[1]

The rapid development of sequencing technologies has led to an exponential increase in the number of known protein sequences, whereas their spatial structures are mostly unresolved. There are currently three widely used methods for the experimental determination of the spatial structure of a protein: X-ray diffraction analysis (XRD),[2] nuclear magnetic resonance spectroscopy (NMR),[3] and cryogenic electron microscopy (cryo-EM).[4] However, it is not always possible to determine the structure of proteins in experiment because of the complexity, high cost, and limitations of experimental techniques. Moreover, not all proteins can be easily obtained in crystalline form, which is necessary for the XRD; a particular problem arises with membrane proteins.[5] In addition, these experimental methods involve complex manipulations with a protein molecule, which can lead to a change in its conformation and, consequently, to artifacts. As a result, although UniProtf8KB/TrEMBL[6] protein database contains more than 200 million known protein sequences, only a tiny fraction of them (slightly more than 150 000 as of March 3, 2021) have their spatial structures experimentally identified at the atomic level and listed in the protein structure bank (PDB).[7] This reflects a large gap between the identification of the sequence of a protein and determination of its structure.

However, we can avoid the complexity and disadvantages of the experimental techniques that slow down the search for three-dimensional structures. The spatial structure of a protein can be, in principle, predicted using theoretical methods based on physical or empirical approximations used for energy-based search for the protein structure,[8] or on its recognition using "big data" and machine learning.[9–11] Such predictions can significantly simplify the first stages of research in molecular biology and medicine.

The energy-based approach which we develop in this work is based on the concept formulated and experimentally confirmed by Anfinsen,[12] a Nobel laureate in chemistry, that all the information needed to fold a protein into its native structure is encoded in its amino acid sequence. Another important application of theoretical methods is the refinement of proteins' crystal structures obtained in experiment: according to the PDB, less than half of them have resolution of better than 2 Å. In addition, these algorithms can enable the construction of such protein sequences that will have a predetermined tertiary structure and, consequently, properties. The solution to this problem is especially important for the pharmaceutical technology and immunology.

Thus, the development of computationally inexpensive and effective theoretical methods for predicting proteins' three-dimensional structures is extremely important.

1.2 | Current methods for predicting the protein structure

According to the thermodynamic hypothesis proposed by Anfinsen[12] and later confirmed by experiments, the native conformation of a protein corresponds to the global minimum of free energy. Using this idea, we can compare different protein conformations and evaluate which one is more stable. Therefore, the central problem of modern algorithms for energy-based predicting the spatial conformation of proteins is finding structures with the minimum potential energy (or rather free energy, to calculate the stability at a finite temperature).[12] Computationally, this is a very difficult task because of a huge number of degrees of freedom that exist in biomolecules. For an ideal prediction, these methods must explore the space of all possible structures, which is astronomically large. This problem, described within the framework of Levinthal's paradox,[13] is the impossibility to enumerate all conformations with exhaustive search. It is worth remembering that there are intrinsically disordered proteins, as well as proteins with multiple stable conformations, but we will not consider them in this article.

Currently, the energy-based protein structure prediction methods are classified by their use of the known experimental structures from the PDB database: template-based (comparative modeling) and template-free (de novo modeling) methods. The template-based methods build protein models by comparing amino acid sequences of a new protein with the sequences of experimentally resolved protein structures and using the experimentally resolved three-dimensional protein structures as templates. The idea that protein chains with similar amino acid sequences will eventually fold into similar three-dimensional conformations follows from the evolutionary proximity of such proteins leading to their very close structural similarity.[14,15] Finding such templates can significantly speed up the prediction of the tertiary structure of a protein.

The template-based methods are often more accurate than the template-free ones, but they are only successful when similar structures are available in the PDB library. In contrast, the template-free methods do not rely on any known protein structures and perform a conformational search using only the amino acid chain of a protein.

Several methods[10,16–22] for the "de novo modeling" have been developed and used to determine protein structures; the most accurate of these methods combine template-based and de novo approaches with the big data-based machine learning.[9–11,18–20] For example, Rosetta,[22] an algorithm developed by D. Baker's laboratory and now used by many scientists around the world, searches for local matches of the amino acid sequences of the original protein with other proteins, then the resulting templates assembled from different structures are used as a starting point for the de novo modeling.

Protein structure prediction methods are strongly dependent on the way the energy is calculated. Potential energies can be accurately estimated using quantum chemistry methods (e.g., density functional theory); however, it is computationally expensive for a system consisting of thousands of atoms, like proteins. Besides, the water that surrounds the protein should be included in the calculations and, among other things, the electrostatic interactions of the protein atoms should be decreased by water by nearly two orders of magnitude. Therefore, to estimate the potential energies of protein structures, empirical force fields are used. In addition, in a living cell, proteins exist at a finite temperature, therefore the entropy contribution should be taken into account. One of the possible solutions is implemented in Rosetta[22]: instead of the potential energy, it uses a "scoring function" which is a proxy to free energy. It consists of the potential energy and statistical terms (which include probabilities in accordance with the Ramachandran map, frequencies of occurrences of rotamers in experimental structures etc.[23]).

Another important technique is molecular dynamics (MD). Nowadays, the state-of-the-art supercomputers can run MD simulations for proteins on millisecond time scale.[24] This very powerful tool helps find not only the stable protein structure, which corresponds to the Gibbs free energy minimum, but also the entire path of molecular folding that led to it. This approach has been applied by various teams around the world, such as the work of D. E. Shaw's group,[25,26] which used Anton supercomputer. The MD approach has also been used in the refinement category of the CASP competitions.[27]

Recently, new machine learning methods became widely used for modeling complex chemical systems. In recent works, deep learning techniques have successfully predicted three-dimensional protein structures.[9–11,20,21,28] Among the machine learning methods, it is important to note the progress made by the DeepMind team in the protein structure prediction. The algorithms they developed, AlphaFold and AlphaFold2,[11] have demonstrated excellent prediction accuracy in the CASP13 and CASP14 competitions. These results give further impetus to solving the problem of protein structure prediction.

However, all the deep learning techniques reported so far use sequence similarity of the protein structures from the PDB and statistical patterns and thus obtain “recognitions” rather than “predictions”. These methods do not focus on describing the physics of the interactions within a protein, which limits their field of application, making it difficult to work with completely new, currently unknown motifs of protein sequences. Although the use of neural networks created a large field for further discussions, it did not finally solve all the problems.

The goal of this work is to develop an energy-based method for predicting the tertiary structure of a protein based on its amino acid sequence, which makes it possible to quickly find stable structures of protein chains in water without the use of homologues. In our work, already published interaction potentials are used, without trying to build an optimal force field. We used the implicit water to simulate the protein environment in our work. Therefore, after calculating energies using the force fields we got a mean-force potential, which takes into account the entropy component of the environment. As an approximation, we considered the entropy of a protein chain in a “solid” protein to be constant. Our method was built on the USPEX evolutionary algorithm,[29–31] which showed high efficiency and reliability in predicting the structures of crystalline solids,[32,33] surfaces,[34] nanoclusters,[35] and molecular crystals.[36] The method does not depend on whether the database contains structures similar to the considered protein. All it needs is the correct parametrization of interactions inside and outside the protein.

An evolutionary algorithm is an optimization method that uses and models the ideas of natural selection to find the “best” solutions (or structures). The best structures are determined by evaluating their fitness function. An example of a fitness function can be energy, volume or any other parameter by which we can compare structures. Further, in our research we will use the term “fitness function” both for mean-force potential energies and scoring functions.

At the beginning, the algorithm creates a set of structures called generation. For each of the structures in the generation, its fitness function is calculated. The best of the resulting structures are then used to produce the next generation by applying variation operators. Other structures with insufficient fitness are discarded so that the “gene pool” of structures is constantly improved. Having performed a set of such calculations, the algorithm iteratively improves the structures and finds the optimum. The USPEX algorithm is a member of the family of evolutionary algorithms and its scheme of work for protein structure prediction is shown in Figure 1.

2 | PRINCIPLES OF THE DEVELOPED ALGORITHM

2.1 | Protein representation

First, it is necessary to define how protein structures are represented in the algorithm. Each atom in a protein is initially defined by three coordinates. To simplify the search space we switch from the coordinate representation for each atom to the torsion angles within and between the adjacent amino acids: φ, ψ, and ω for the main chain and χ for the side chains.[1,33,37] The purpose of the evolutionary algorithm is to find the optimal set of dihedral angles. After the algorithm fixes a certain set of angles, we return from the representation of torsion angles to the coordinate space. In it, we run local optimization using force fields and determine the fitness function of structures. In this work, we considered all the atoms in the side chains. The transition from torsion angles to coordinate space uses the distances between atoms from the Tinker program.

After local optimization, we return to the space of angles φ and ψ and change them using variation operators. The angles ω and χ can change during the relaxation of the structure via gradient descent; no separate variation operator is assumed for them; χ-angles are initialized with the values given in the parameter files of Tinker[38] software. At the beginning the angles φ and ψ are initialized with the values taken randomly from the database and the angles ω are assumed to be 180°: at this value, a deep minimum on the potential energy exists because of sp^2 hybridization of the C—C bond in the amino acid chains.[1] All amino acid residues in proteins, except for some of proline residues, are known to have $\omega \approx 180^\circ$; but 5%–10% of prolines have $\omega \approx 0^\circ$ (this is another, but higher energy minimum for the sp^2 hybridization of the C—C bond[1]). In this work, we, for simplicity, will only consider the proteins where all the prolines have $\omega \approx 180^\circ$ (i.e., have trans- conformations).

2.2 | Diagram of the USPEX algorithm of protein structure prediction

Each protein is defined as a set of pairs of torsion angles. The evolutionary algorithm optimizes the potential energy (or scoring function) as a function of these angles, which we use to compare the resulting structures with each other. To calculate the structure fitness function,

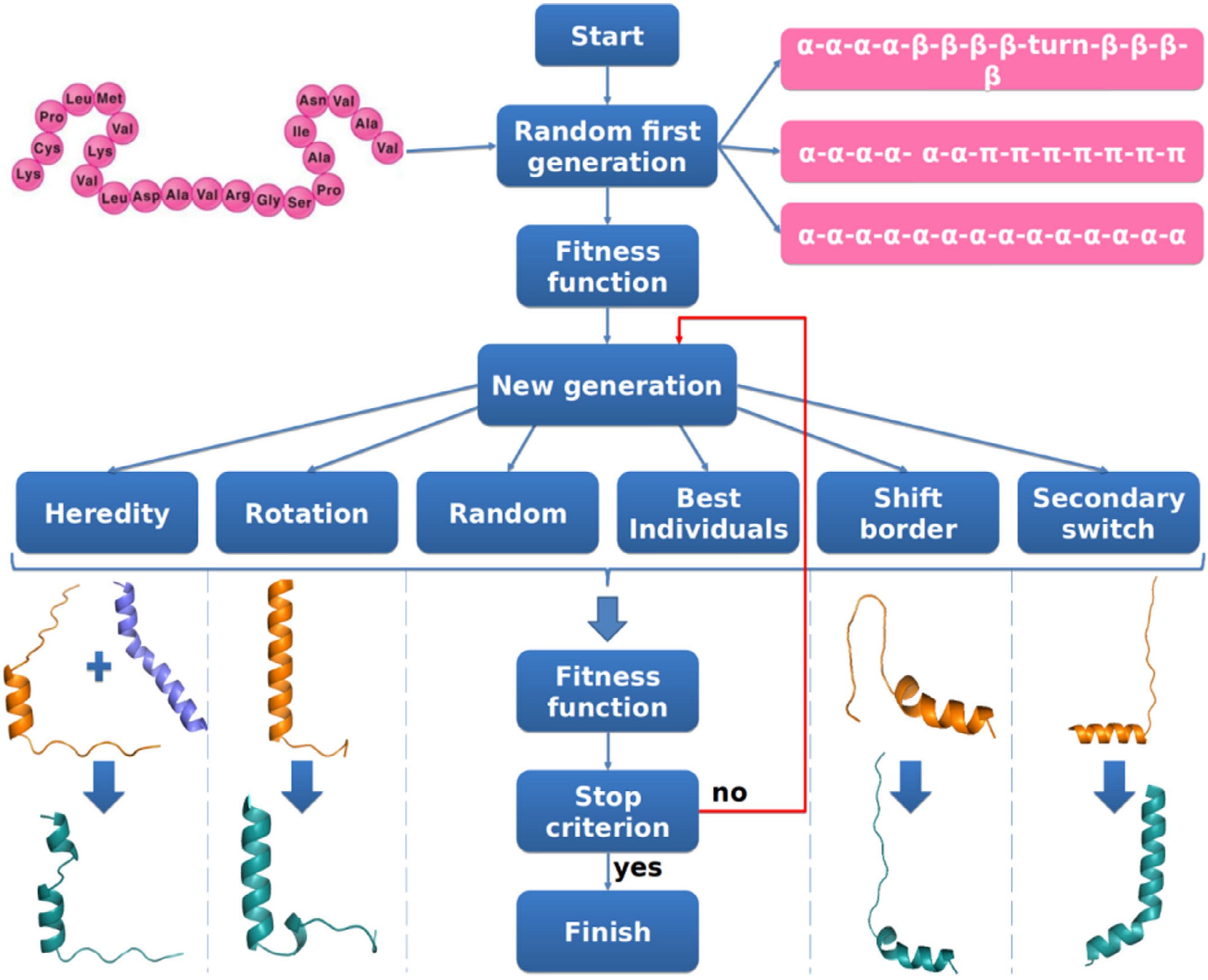
Start
Random first generation
α-α-α-α-β-β-β-β-turn-β-β-β-β
α-α-α-α- α-α-π-π-π-π-π-π-π
α-α-α-α-α-α-α-α-α-α-α-α-α-α
Lys Cys Pro Leu Met Val Lys Val Leu Asp Ala Val Arg Gly Ser Pro Ala Ile Asn Val Ala Val
Fitness function
New generation
Heredity
Rotation
Random
Best Individuals
Shift border
Secondary switch
Fitness function
Stop criterion
no
yes
Finish

FIGURE 1 Scheme of the USPEX algorithm for predicting protein structure. Symbols "α", "β", "turn" denote main-chain conformations of amino acid residues. See explanations in the text.

we used four different force fields often used in literature: Amber ff99SB,[39,40] CHARMM22/CMAP,[41] and OPLS-AA/L[42] from Tinker and REF2015 from Rosetta.[22] The Lazaridis–Karplus[43] and GBSA[44] implicit water models were used for calculations with Rosetta and Tinker, respectively. The details of the selection of potentials and their parameters are presented in Methods section (5.1. Force field parameters and water models).

The developed algorithm (Figure 1) is similar to the earlier used[24–26] USPEX evolutionary algorithm; the main differences come from the random generation of structures and variation operators that work with the heredity of conformations (Heredity), rotation of angles (Rotation), displacement of the boundaries of the secondary structures (ShiftBorder), and replacement of the secondary structures (SecSwitch). «Random» and «Best Individuals» are needed to randomly generate new structures and to preserve the best structures from generation to generation, respectively.

1. Random: The structure generation algorithm splits a protein into several parts, then for each of them randomly takes the main chain angles from one of almost 80 thousand proteins (ranging in length from 8 to 3660 amino acid residues) from the database (https://zhanglab.ccmb.med.umich.edu/library/). Thus, the angles in the resulting structure are not completely random, which helps to avoid generating a set of unrealistic angles.

   After a generation of structures is produced and relaxed, the structures are ranked by fitness. The fittest 70% structures are given probabilities to become parents for the next generation.
2. Heredity: The algorithm randomly selects 25%–75% of the pairs of main chain angles from two parents from the previous generation and combines these parts into a new protein of the same length.
3. Rotation: Up to 20% of sequential main chain angle pairs in a protein are randomly changed.
4. Shift border: A protein is divided into secondary structure segments (i.e. segments where all the residues belong to the same

secondary structures) using STRIDE[45] code STRIDE determines the secondary structures to which each of the amino acids in the protein corresponds. Then, a region with one of the secondary structures is randomly selected. The values of the angles of this secondary structure are assigned to the adjacent amino acids either to the right or to the left of the selected region. The shift of the secondary structure occurs randomly, shifting the border between secondary structure domains by 1–5 amino acid residues.

5. Secondary switch: As above, the protein is again divided into secondary structure segments (i.e., segments where all the residues belong to the same secondary structures) using STRIDE.[47] The values of the angles corresponding to one randomly selected part are replaced with the values corresponding to a different secondary structure type taken randomly from the experimental database.
6. Best individuals: A certain number of the best structures from the previous generation are preserved and passed on to the next generation, and participate in producing offspring again.

2.3 | Computational parameters

2.3.1 | Parameters of population

For optimal operation of the evolutionary algorithm, it is important to determine its internal parameters. To do this, a series of calculations with a different sets of parameters were carried out in each of the tested force fields. It was determined that the population of 250 structures and 100 generations is sufficient for most calculations. These values can be used as defaults for proteins that have up to ~100 amino acids.

2.3.2 | Variation operators

It is important that the structures obtained using variation operators have low energies and are sufficiently diverse, in order to explore conformation space more fully. Operators which satisfy these conditions to a greater extent will be applied more frequently.

To check how different computed structures are, we calculated a «fingerprint»[46] for each of them. In our work, the fingerprint is a way to describe how compact a protein is. For each of the atoms in the structure, a set of spheres (127 in our case) of a fixed radius centered in the coordinates of the atom is constructed. In this consideration, each atom has a finite volume, that is, is not a point particle. The algorithm then calculates how many atoms are on the surface of each sphere, the results are summed up over all the atoms in the structure and normalized. A vector called a fingerprint is thus obtained, the number of elements of which is equal to the number of spheres. The value of each element in the vector is the calculated number of atoms on the surface of a sphere of a certain radius. Such vectors (fingerprints) were calculated for each of the protein chains we have obtained. To visualize the results, we applied the principal component analysis (PCA) to the calculated fingerprints. Taking a linear combination of the vector components describing the structure in a multidimensional space, we switched to a two-dimensional space, where the basis is defined by two principal components. After this transformation, the basis vectors no longer carry any specific physical meaning, but they can describe the calculated structures, retaining most of their variation (in our case, 60%). This way, in the two-dimensional space, we can visualize the distribution of protein structures created using different variation operators (Figure 2).

These calculations were carried out for each of the proteins, followed by a visual analysis of the result. Figure 2A shows that the

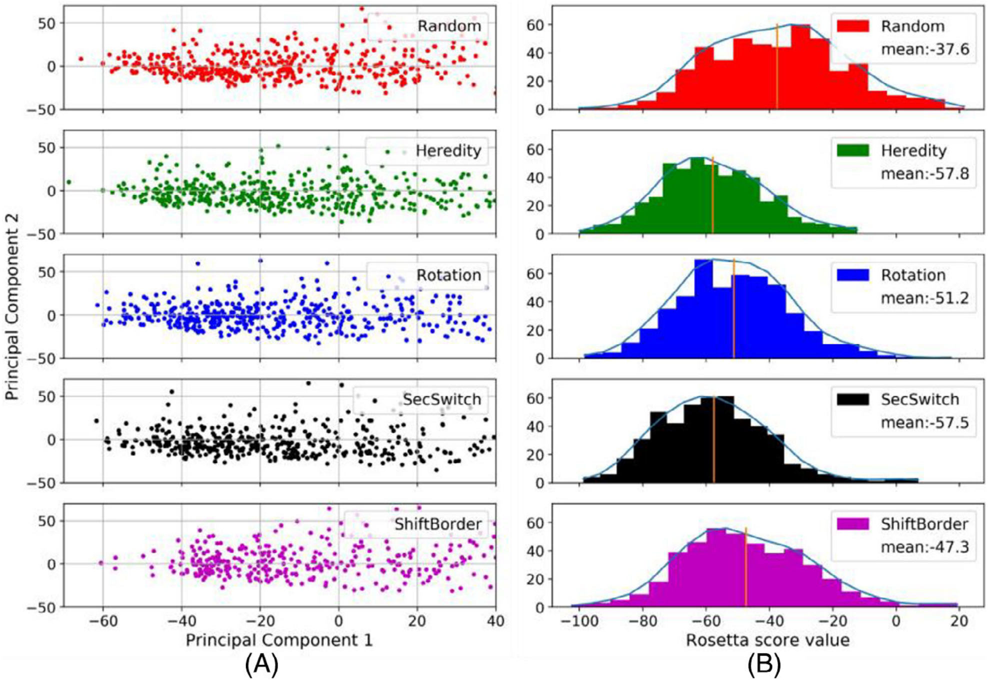

FIGURE 2 (A) Distribution of the generated structures for protein 1shf after the PCA analysis, created using different variation operators (with subsequent pseudo-energy minimization) during the USPEX run (red—Random, green—Heredity, blue—Rotation, black—SecSwitch, purple—ShiftBorder), in the coordinates of the principal components. (B) Pseudo-energy ("scoring function") distribution of the generated protein structures. An orange line marks the average pseudo-energy of the distribution. The ordinate represents the number of structures.

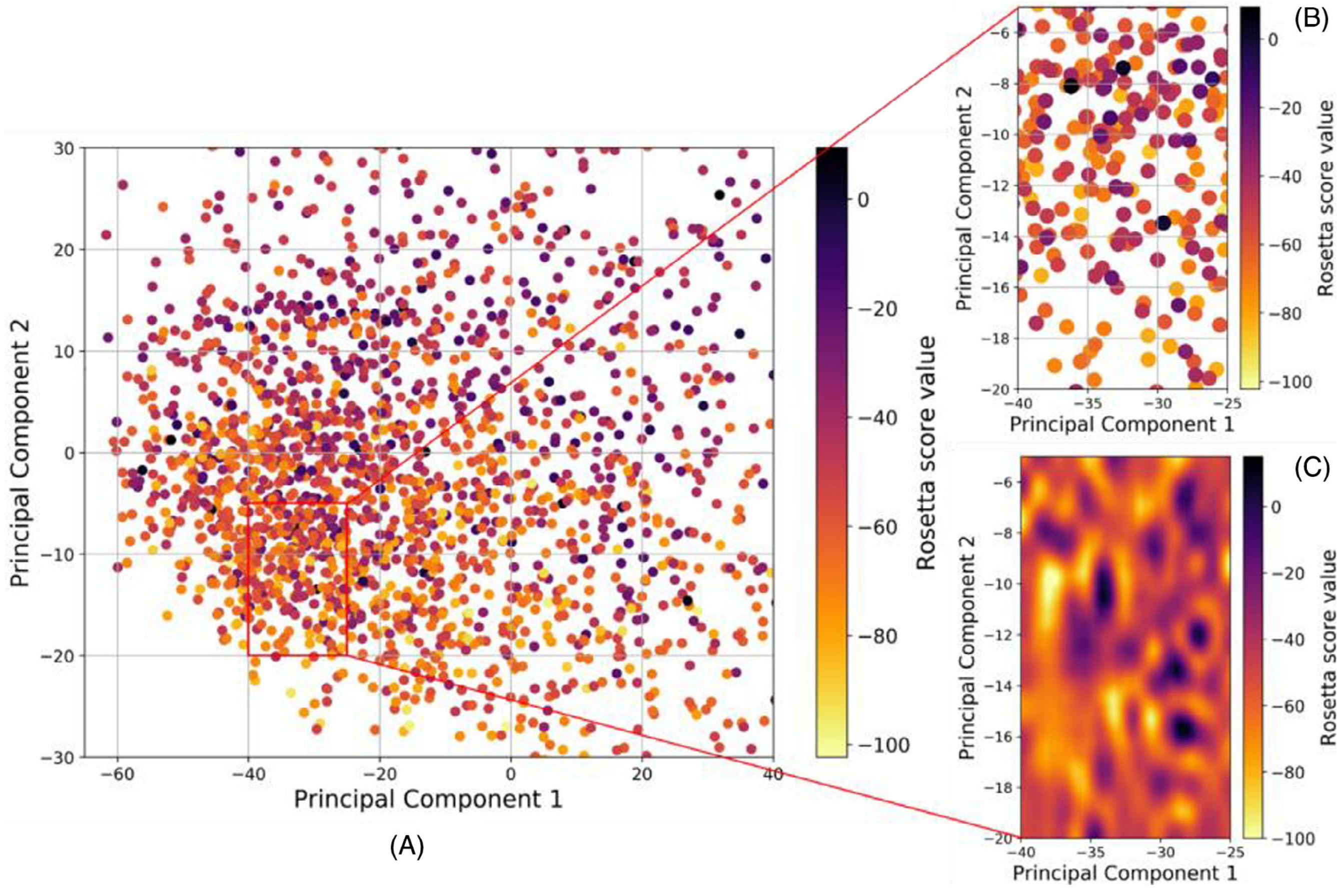

FIGURE 3 (A) Fitness function surface of the generated structures for the chain of protein 1shf in the principal component basis. (B) Magnified area of the fitness function. (C) Interpolation of the fitness function surface presented in Figure 3B.

structures obtained using different variation operators are equally distributed over the space of the PCA features, only the result of ShiftBorder operator is slightly different: there are very few structures obtained using ShiftBorder operator on the left side of the distribution in coordinates of the principal components (Figure 2A). In addition, using visual analysis of the energy distribution of the structures, one notes that all variation operators result in energies that are on average lower than in random structures, while among the variation operators, the ShiftBorder produces the highest average energy (Figure 2B). This trend has been observed for all proteins studied here. This gives a reason to reducing the number of structures obtained using ShiftBorder operator in favor of other operators. Based on such analysis, we chose the following ratios of variation operators to be used initially to create structures:

30%—Generation of a new random structure,
25%—Heredity,
20%—Rotation,
15%—Secondary switch,
10%—ShiftBorder.

To generate low-energy structures more often, these ratios are dynamically evolved.[47] Comparing the energy distributions of structures (Figure 2B), we conclude that the use of variation operators is justified because it shifts the distribution of structures closer to the low-energy region (compared with random structure generator).

To find out whether the generated structures are evenly distributed or concentrated in one area in the space of principal components, we visualized the fitness function surface of the protein 1shf (Figure 3).

The structures with the Rosetta scoring function below 20 units are presented in Figure 3. It shows that the algorithm explores not

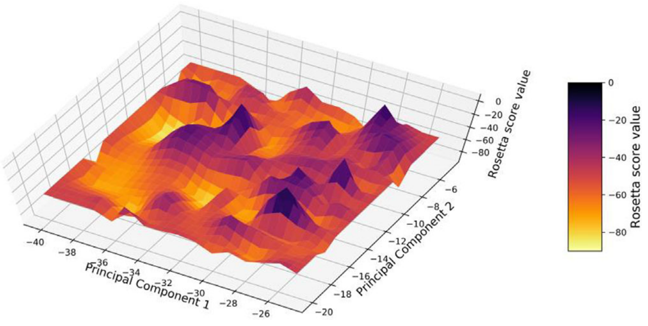

FIGURE 4 3D visualization of the fitness function surface of the protein 1shf in the principal component basis.

TABLE 1 Fitness functions of the best structures predicted in five different USPEX runs with different force fields compared with the values obtained for the experimental protein structure using the same potentials. The last column shows the results obtained using Rosetta Abinitio. The root-mean-square deviation (RMSD) of the positions of the C_α atoms from their positions in the real structure as well as GDT_TS score is shown after the energies. The minimum energies obtained in five runs are set in bold. The results with the lowest RMSD and highest GDT scores are boxed for each protein. The energies of the protein structures obtained from the PDB are underlined.

Protein \force field	Amber ff99SB	CHARMM22/CMAP	OPLS-AA/L	REF2015	Rosetta Abinitio (validation)
	$\frac{kcal}{mol}$|Å|GDT	$\frac{kcal}{mol}$|Å|GDT	$\frac{kcal}{mol}$|Å|GDT	Ros_score|Å|GDT	Ros_score|Å|GDT
2rvd	$E_{exp} = -427.2$	$E_{exp} = -320.8$	$E_{exp} = -644.3$	$E_{exp} = -15.6$	N/A (too short chain)
	$E_1 = -423.9$|0.8|97.5	$E_1 = -$**320.7**|0.6|97.5	$E_1 = -646.4$|2.6|80.0	$E_1 = -24.3$|0.7|97.5	
	$E_2 = -425.8$|0.6|100.0	$E_2 = -320.5$|0.6|97.5	$E_2 = -646.7$|1.0|95.0	$E_2 = -$**28.3**|1.8|85.0	
	$E_3 = -424.7$|0.5|100.0	$E_3 = -320.5$|0.6|97.5	$E_3 = -$**647.4**|2.7|77.5	$E_3 = -24.3$|0.7|97.5	
	$E_4 = -$**425.9**|0.5|100.0	$E_4 = -$**320.7**|0.6|97.5	$E_4 = -646.8$|3.0|80.0	$E_4 = -24.7$|0.9|95.0	
	$E_5 = -$**425.9**|0.7|97.5	$E_5 = -320.5$|0.6|97.5	$E_5 = -$**647.4**|2.7|77.5	$E_5 = -24.3$|0.7|97.5	
2jof	$E_{exp} = -653.6$	$E_{exp} = -574.3$	$E_{exp} = -1007.4$	$E_{exp} = -54.0$	NA (too short chain)
	$E_1 = -667.9$|1.9|88.8	$E_1 = -575.3$|4.2|63.8	$E_1 = -1023.7$|6.1|48.8	$E_1 = -51.3$|1.8|85.0	
	$E_2 = -662.9$|2.5|78.8	$E_2 = -582.3$|1.2|92.5	$E_2 = -1021.1$|5.7|57.5	$E_2 = -$**52.7**|3.9|65.0	
	$E_3 = -662.0$|5.0|65.0	$E_3 = -$**584.1**|2.4|87.5	$E_3 = -1020.9$|4.2|58.8	$E_3 = -50.9$|4.7|62.5	
	$E_4 = -666.7$|5.2|62.5	$E_4 = -578.2$|3.9|66.3	$E_4 = -1021.3$|5.1|56.3	$E_4 = -50.6$|4.7|61.3	
	$E_5 = -$**668.3**|1.8|88.8	$E_5 = -578.2$|3.9|65.0	$E_5 = -$**1025.4**|4.6|55.0	$E_5 = -49.7$|5.4|60.0	
1fme	$E_{exp} = -1782.5$	$E_{exp} = -2066.98$	$E_{exp} = -2148.8$	$E_{exp} = -51.6$	-69.1|6.8|65.2
	$E_1 = -1814.3$|8.0|46.4	$E_1 = -2115.1$|8.3|41.0	$E_1 = -2195.7$|6.6|55.4	$E_1 = -67.7$|7.9|54.5	
	$E_2 = -1817.6$|8.1|49.1	$E_2 = -$**2116.5**|7.3|54.5	$E_2 = -2191.1$|7.3|47.3	$E_2 = -$**72.9**|6.8|66.1	
	$E_3 = -$**1826.1**|7.9|53.6	$E_3 = -2113.4$|8.1|44.6	$E_3 = -2186.4$|6.3|59.8	$E_3 = -68.5$|7.4|59.8	
	$E_4 = -1822.1$|7.4|54.5	$E_4 = -2113.4$|7.7|50.0	$E_4 = -2205.4$|6.7|42.9	$E_4 = -65.6$|7.3|62.5	
	$E_5 = -1818.9$|8.0|52.7	$E_5 = -2109.7$|6.5|47.3	$E_5 = -$**2210.4**|6.7|46.4	$E_5 = -70.8$|3.0|75.9	
1enh	$E_{exp} = -3543.0$	$E_{exp} = -3820.9$	$E_{exp} = -3578.7$	$E_{exp} = -159.1$	-155.3|3.6|72.2
	$E_1 = -3508.9$|10.5|37.0	$E_1 = -$**3822.3**|11.2|41.2	$E_1 = -3555.8$|9.3|37.0	$E_1 = -140.5$|8.9|40.7	
	$E_2 = -3508.6$|9.2|53.7	$E_2 = -3812.4$|14.1|37.5	$E_2 = -3563.4$|9.6|35.6	$E_2 = -126.1$|17.1|34.7	
	$E_3 = -3509.6$|9.9|37.5	$E_3 = -3812.4$|13.2|38.4	$E_3 = -3595.3$|11.6|34.3	$E_3 = -$**142.1**|9.1|40.7	
	$E_4 = -$**3513.7**|10.3|27.8	$E_4 = -3817.3$|18.0|31.9	$E_4 = -3582.9$|10.6|35.6	$E_4 = -141.1$|11.7|38.4	
	$E_5 = -3507.3$|11.4|45.4	$E_5 = -3817.8$|12.8|38.9	$E_5 = -$**3596.7**|10.0|31.0	$E_5 = -135.0$|10.9|38.9	
1shf	$E_{exp} = -2239.8$	$E_{exp} = -1964.2$	$E_{exp} = -3446.6$	$E_{exp} = -194.5$	-152.7|11.0|46.6
	$E_1 = -2209.9$|13.6|21.2	$E_1 = -2010.1$|18.1|17.8	$E_1 = -3424.0$|11.3|22.5	$E_1 = -133.2$|13.0|20.3	
	$E_2 = -2209.0$|13.1|20.3	$E_2 = -2009.4$|18.5|18.2	$E_2 = -3430.7$|12.7|23.3	$E_2 = -146.3$|11.7|23.3	
	$E_3 = -2212.7$|11.6|23.3	$E_3 = -2011.0$|18.0|19.1	$E_3 = -$**3442.6**|12.1|23.7	$E_3 = -138.6$|14.6|19.5	
	$E_4 = -$**2220.3**|11.8|22.5	$E_4 = -$**2016.5**|15.9|19.1	$E_4 = -3424.3$|12.6|22.9	$E_4 = -145.8$|13.2|22.9	
	$E_5 = -2211.4$|11.4|26.3	$E_5 = -2014.2$|17.9|18.2	$E_5 = -3422.8$|10.7|20.7	$E_5 = -$**147.7**|10.9|23.7	
2a3d	$E_{exp} = -3009.1$	$E_{exp} = -3139.8$	$E_{exp} = -4548.1$	$E_{exp} = -185.0$	-225.7|2.8|72.9
	$E_1 = -3062.8$|11.9|40.8	$E_1 = -3222.8$|15.3|44.5	$E_1 = -$**4615.0**|13.7|21.2	$E_1 = -212.6$|4.1|58.6	
	$E_2 = -3068.0$|15.7|34.6	$E_2 = -3221.8$|13.2|34.2	$E_2 = -4599.5$|16.1|26.0	$E_2 = -$**214.2**|9.7|43.2	
	$E_3 = -3071.7$|12.2|39.7	$E_3 = -3219.6$|15.4|32.5	$E_3 = -4597.2$|12.1|34.9	$E_3 = -209.0$|4.2|64.4	
	$E_4 = -$**3071.8**|7.7|47.9	$E_4 = -3221.7$|6.3|43.8	$E_4 = -4602.7$|10.6|34.2	$E_4 = -212.1$|9.1|45.5	
	$E_5 = -3067.7$|13.4|35.3	$E_5 = -$**3228.6**|13.3|33.9	$E_5 = -4611.7$|16.3|28.1	$E_5 = -212.8$|5.9|50.0	
1cei	$E_{exp} = -3292.4$	$E_{exp} = -3267.9$	$E_{exp} = -5406.1$	$E_{exp} = -265.3$	-221.5|12.1|29.4
	$E_1 = -3295.4$|12.5|27.4	$E_1 = -3307.8$|20.3|27.3	$E_1 = -5412.7$|13.6|22.9	$E_1 = -212.2$|16.7|29.4	
	$E_2 = -3271.8$|13.8|26.5	$E_2 = -3306.8$|18.0|26.8	$E_2 = -$**5417.4**|12.8|27.1	$E_2 = -222.9$|11.6|30.3	
	$E_3 = -3273.4$|14.5|27.1	$E_3 = -$**3312.7**|16.6|26.8	$E_3 = -5414.4$|16.2|21.2	$E_3 = -$**227.7**|12.9|27.4	
	$E_4 = -$**3296.0**|10.0|37.4	$E_4 = -3302.1$|23.7|25.0	$E_4 = -5401.1$|16.3|22.6	$E_4 = -223.5$|10.9|27.9	
	$E_5 = -3289.6$|12.3|28.8	$E_5 = -3303.1$|14.7|27.6	$E_5 = -5411.6$|18.3	$E_5 = -211.4$|17.9|25.0	

only the global minima, but also many different isolated local minima, which are separated by higher energy barriers. Inset (B) is given to illustrate that even in a small area of space deep local minima separated by high energy barriers are presented nearby. In addition to many "favorable" structures with low energy found in this region, there are also high-energy conformations. The potential energy surface of proteins is extremely complex, rough, and has many local minima (Figure 3C and Figure 4), and must significantly complicate the search for the real structure. Figure 4 is a 3D rendering of the area shown in Figure 3A.

To avoid local minima,[35] it may be useful to use the antiseeds technique earlier developed within USPEX method.[29–31] This mechanism adds a penalty—a cumulative positive term—to the fitness function of those structures that have already been sampled during the calculation, and structures with similar conformations, forcing the algorithm to sample other areas of configurational space.

2.4 | Validation of results

Regarding the accuracy and required computational resources, we compared the present extension of the USPEX method for protein structure prediction with the Rosetta Abinitio protocol (See Section 5). Note that Rosetta web service can only create fragment files for proteins with more than 27 residues.

After each run, structures with the lowest fitness function were selected and visualized using PyMOL[48] software. To assess the quality of protein prediction, the following metrics were used: the proximity of the fitness function of the calculated structure to the fitness function of the experimental one, GDT_TS between the optimally superimposed structures, and RMSD, where RMSD is:

$$\mathrm{RMSD} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} d_i^2}, \qquad (1)$$

where n is the number of pairs of equivalent C_α atoms in proteins and d_i is the distance between the atoms in the i-th pair of the C_α atoms.

GDT_TS is better at detecting similarities in fold than RMSD. GDT_TS is computed using alpha carbon atoms and is reported as a percentage range from 0 (a meaningless prediction) to 100 (a perfect prediction). It is useful to highlight the following characteristic values: random prediction gives about 20 GDT score. Correct determination of the gross topology gives about 50 GDT score and accurate topology around 70 GDT score. With correct prediction of all C_α atoms, GDT_TS goes up to 100.[49]

The set of cut-off distances that was used to calculate the GDT_TS is 1, 2, 4, and 8 angstroms.

3 | RESULTS

Our protein structure prediction method was tested on seven different peptides and proteins: 2rvd (10 amino acids), 2jof (20 aa), 1fme (28 aa), 1enh (54 aa), 1shf (59 aa), 2a3d (73 aa), and 1cei (85 aa). The prediction results are arranged by the length of the amino acid sequence, and hence the complexity of prediction. To validate the results and obtain statistics, five USPEX runs were carried out for each protein chain (Table 1, Figure 5).

Additional comments detailing how the algorithm works for each protein are given in the Supplementary Information (8.4. Description of results). Visualizations of the best models of each of the considered proteins with each potential are given in the Supplementary Information (Figures S7–S13).

In general, even when USPEX does not find a global minimum, its final structures look like folded globules. This shows that these structures can indeed be locally stable conformations of a protein or its transitional forms.

During the USPEX calculations, we often obtained conformations with fitness functions lower than those of the experimental structures, which is incompatible with Anfinsen's dogma (that proteins adopt thermodynamically stable conformations). However, numerous experiments assure us that Anfinsen's dogma is correct. Therefore, one may conclude that currently available force fields are insufficiently accurate.

It is worth remembering that some proteins begin to fold even before the entire chain leaves the ribosome, and the pathways leading to the final state may differ significantly for in silico and in vivo.[50] However, we work with relatively small proteins and peptides and can expect our proteins to fold into a conformation that corresponds to the minimum free energy, regardless of the conditions of the experiment.

In addition, note that our physics-based protein structure prediction searches were conducted in the absence of other proteins, including chaperones, with which the protein can interact during the folding process and which can influence the final state.

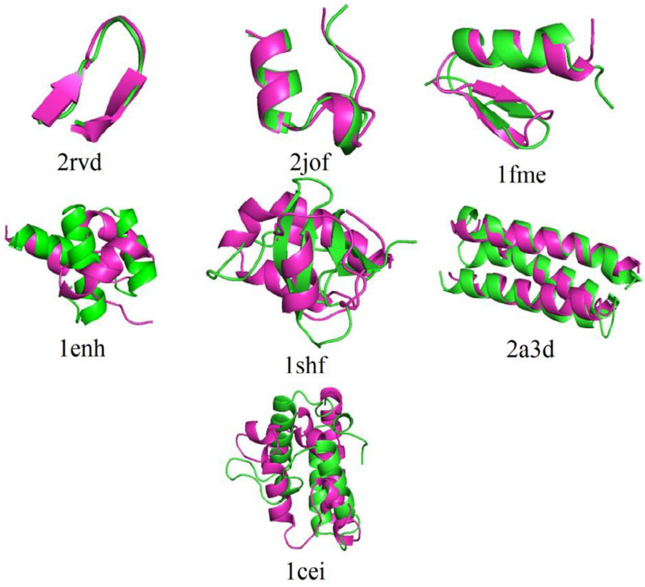

FIGURE 5 Visual comparison of the best (in terms of the RMSD) protein structures predicted using the USPEX algorithm (purple) with the real conformations from the PDB, shown in green.

4 | CONCLUSION

This study presents an extension of the evolutionary algorithm USPEX aimed at predicting the tertiary structure of a protein on the basis of its amino acid sequence. In this method, the optimization algorithm searches for a set of main chain dihedral angles that matches the structure with the minimum fitness function. After the algorithm fixes some set of angles, we go to the full-atomic representation and the force fields presented in Tinker and Rosetta are used for structure relaxation and energy calculation. To avoid trapping in local minima, we create new structures using several physically motivated variation operators: Heredity, Rotation, SecSwitch, and ShiftBorder.

To test the developed method, we predicted stable structures of various proteins ranging in length from 10 to 85 amino acid residues and compared them with the results obtained using Rosetta Abinitio. In two out of five cases, our evolutionary method showed greater accuracy; in other three runs it obtained structures with slightly higher but similar fitness function. We found structures close to the real ones for the proteins 2rvd, 2jof, 1fme, and 2a3d. REF2015 force field, on average, showed better results as it is a proxy to free energy and can be singled out for further work.

Most of the protein structures obtained using the developed algorithm have conformations with fitness functions lower than that of the real structure. On the one hand, this suggests that our method is very successful in searching for very deep minima. On the other hand, this also suggests that fitness function surfaces of these proteins are different from the exact free energy surfaces where the native structures reside, have different global minima, and are not sufficiently accurate for protein structure prediction. The experimental evidence suggest that proteins do adopt globally optimal structures and currently available force fields[51] are too crude for reliable protein structure prediction. Their further development should be a very significant step for the computational prediction of protein structures.

5 | METHODS

5.1 | Force field parameters and water models

To relax protein structures, we used Tinker[38] and Rosetta[22] packages, which include force fields that differ in the parameterization of the interactions of individual parts of a protein (Equation S1).

The choice of a force field has a significant impact on final results. There are 17 different sets of potentials included in Tinker and one potential set in Rosetta—REF2015.[22] For each potential–potential pair, a single relaxation of the same set of structures was carried out and the correlation between the energies obtained during the relaxation was analyzed (Figure S1). On the basis of this analysis, we chose four force fields, which are often used in other works: Amber ff99SB,[39,40] CHARMM22/CMAP,[41] OPLS-AA/L,[42] and REF2015.

To relax protein structures, Tinker and Rosetta use the gradient descent method. The criterion for stopping the calculations is the absolute value of the energy gradient. In calculations using Tinker (ver. 8.7.2), the RMS gradient value of 0.01 kcal/mol/Å was used. For Rosetta (ver. 2020.08) calculations, the accuracy of minimization within the FastRelax protocol was 0.00001 in scoring function units. To visualize the resulting structures, PyMOL[48] software was used.

The interaction of a protein with its environment is also quite important. In a real cell, a protein can react with a membrane, other proteins, and a solution. Most of these interactions are impossible to model due to our limited knowledge about them. To model interactions with water, two approaches can be used. In the explicit water models, the environmental molecules are explicitly created in a simulated volume using periodic boundary conditions. Then, during the relaxation, the forces (and velocities in the case of dynamics) are recalculated at each step for each atom, including those of waters. However, modeling all interactions and recalculating forces greatly slows down the minimization. No such drawback exists in the implicit water models, which use a certain potential of mean force (PMF) to describe interactions with water. They are much faster, although slightly less accurate. In our work, we used Lazaridis–Karplus[43] implicit water model with REF2015 force field and GBSA[44] implicit water model with force fields of the Tinker software package.

5.2 | Rosetta Abinitio parameters

Structure prediction using Rosetta Abinitio is based on a Monte Carlo algorithm and fragment files, which are sets of template angles for different combinations of amino acid sequences. Fragment files were generated by Rosetta web server using the "Exclude Homologues" protocol. The angles taken from these templates are inserted into those regions of a protein that contain the amino acid residue sequences identical to the template. After each insertion, the increase in energy is estimated, then the structure change is applied or rejected according to the Metropolis algorithm[52] using the effective temperature. The kinetic energy of proteins is not explicitly taken into account in this approach, because within classical mechanics it is a structure-independent constant at a given temperature.

The prediction of Rosetta Abinitio totally depends on whether there are templates in the protein database whose angles coincide with the angles of the protein considered. Our algorithm does not depend on these parameters and, despite using databases of protein structures, does not check the coincidence of amino acid residues in different proteins. Its work is based entirely on variation operators and our evolutionary algorithm.

Due to the random nature of Monte Carlo algorithms, different runs of Rosetta Abinitio can lead to different local minima. Therefore, many independent runs of this algorithm are carried out, and the structure with the lowest scoring function is considered the most stable one.

As we mentioned earlier, 25 000 protein chain structures were generated on average in each run of the USPEX algorithm (100 generations of 250 structures each), and for each protein only 5 runs were performed. To compare and evaluate our results, 25 000 runs were performed in Rosetta Abinitio.

AUTHOR CONTRIBUTIONS

Pavel Rachitskii: Investigation; writing – original draft; conceptualization; visualization; writing – review and editing; validation. **Ivan Kruglov:** Conceptualization; investigation; writing – review and editing; methodology. **Alexei V. Finkelstein:** Writing – review and editing. **Artem R. Oganov:** Supervision; writing – review and editing; methodology; conceptualization.

ACKNOWLEDGMENTS

Artem R. Oganov thanks M. Rakitin for his help at early stages of this project, A. Masunov for advice on the selection of force fields, and I. Yakutenko for discussions which have resulted in the kinetic protein folding algorithm. Artem R. Oganov acknowledges support from Russian Science Foundation (grant 19-72-30043). Alexei V. Finkelstein acknowledges funding from the Russian Science Foundation (grant 21-14-00268). I. Kruglov acknowledges support from the Ministry of Science and Higher Education of the Russian Federation (FSMG-2021-0005). We also thank RosettaCommons for making Rosetta freely available for academic use.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/prot.26478.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are currently in the development branch of USPEX code https://uspex-team.org, and can be provided upon request.

ORCID

Pavel Rachitskii https://orcid.org/0000-0003-0533-7940

REFERENCES

1. Finkelstein AV, Ptitsyn OB. *Protein Physics. A Course of Lectures*. 2nd ed., Academic Press, An Imprint of Elsevier Science; 2016.
2. Drenth J. *Principles of Protein X-Ray Crystallography*. Springer; 2007.
3. Bothwell JHF, Griffin JL. An introduction to biological nuclear magnetic resonance spectroscopy. *Biol Rev*. 2011;86:493-510.
4. Wang H-W, Wang J-W. How cryo-electron microscopy and X-ray crystallography complement each other. *Protein Science*. 2016;26(1):32-39. Portico. https://doi.org/10.1002/pro.3022
5. Birch J, Cheruvara H, Gamage N, Harrison JP, Lithgo R, Quigley A. Changes in membrane protein structural biology. *Biology*. 2020;9:401.
6. Hancock JM, Zvelebil MJ, Zvelebil MJ. UniProt. *Dictionary of Bioinformatics and Computational Biology*; Wiley-Liss; 2004.
7. Protein Data Bank. *RCSB PDB: Homepage*. Rcsb Pdb; 2019.
8. Dodson EJ. Protein predictions. *Nature*. 2007;450:176-177.
9. Senior AW, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning. *Nature*. 2020;577:706-710.
10. Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D. Improved protein structure prediction using predicted interresidue orientations. *Proc Natl Acad Sci U S A*. 2020;117:1496-1503.
11. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596:583-589.
12. Anfinsen CB. Principles that govern the folding of protein chains. *Science*. 1973;181:223-230.
13. Levinthal C. How to fold graciously. *Mössbauer Spectrosc Biol Syst Proc*. 1969;41:22-24.
14. Lesk AM, Chothia C. The response of protein structures to amino acid sequence changes. *Philos Trans R Soc London*. 1986;A317:345-356.
15. Chothia C. One thousand families for the molecular biologist. *Nature*. 1992;357:543-544.
16. Xu D, Zhang Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins Struct. Funct. Bioinforma*. 2012;80:1715-1735.
17. Bhattacharya D, Cao R, Cheng J. UniCon3D: De novo protein structure prediction using united-residue conformational search via stepwise, probabilistic sampling. *Bioinformatics*. 2016;32:2791-2799.
18. Gao P, Wang S, Lv J, Wang Y, Ma Y. A database assisted protein structure prediction method: via a swarm intelligence algorithm. *RSC Adv*. 2017;7:39869-39876.
19. Bhattacharya D, De Cheng J. Novo protein conformational sampling using a probabilistic graphical model. *Sci Rep*. 2015;5:16332.
20. Evans R, Jumper J, Kirkpatrick J, et al. De novo structure prediction with deep-learning based scoring. *Thirteen Crit Assess Tech Protein Struct*. 2018;77:6.
21. Qin Z, Wu L, Sun H, et al. Artificial intelligence method to design and fold alpha-helical structural proteins from the primary amino acid sequence. *Extrem Mech Lett*. 2020;36:100652.
22. Alford RF, Leaver-Fay A, Jeliazkov JR, et al. The Rosetta all-atom energy function for macromolecular modeling and design. *J Chem Theory Comput*. 2017;13:3031-3048.
23. Gooch, J. W. Ramachandran*Encyclopedic Dictionary of Polymers, Springer*; 2011.
24. Brini E, Simmerling C, Dill K. Protein storytelling through physics. *Science*. 2020;370:6520.
25. Lindorff-Larsen K, Piana S, Dror RO, Shaw DE. How fast-folding proteins fold. *Science*. 2011;80(334):517-520.
26. Robustelli P, Piana S, Shaw DE. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc Natl Acad Sci U S A*. 2018;115:E4758-E4766.
27. Terashi G, Kihara D. Protein structure model refinement in CASP12 using short and long molecular dynamics simulations in implicit solvent. *Proteins Struct Funct Bioinforma*. 2018;86:189-201.
28. AlQuraishi M. End-to-end differentiable learning of protein structure. *Cell Syst*. 2019;8:292-301.
29. Oganov AR, Glass CW. Crystal structure prediction using ab initio evolutionary techniques: principles and applications. *J Chem Phys*. 2006;124:244704.
30. Oganov AR, Lyakhov AO, Valle M. How evolutionary crystal structure prediction works-and why. *Acc Chem Res*. 2011;44:227-237.
31. Lyakhov AO, Oganov AR, Stokes HT, Zhu Q. New developments in evolutionary structure prediction algorithm USPEX. *Comput Phys Commun*. 2013;184:1172-1182.
32. Dong X, Oganov AR, Goncharov AF, et al. A stable compound of helium and sodium at high pressure. *Nat Chem*. 2017;9:440-445.
33. Ma Y, Eremets M, Oganov AR, et al. Transparent dense sodium. *Nature*. 2009;458:182-185.
34. Tikhomirova KA, Tantardini C, Sukhanova EV, et al. Exotic two-dimensional structure: the first case of hexagonal NaCl. *J. Phys. Chem. Lett*. 2020;11:3821-3827.
35. Lepeshkin SV, Baturin VS, Uspenskii YA, Oganov AR. Method for simultaneous prediction of atomic structure and stability of nanoclusters in a wide area of compositions. *J Phys Chem Lett*. 2019;10:102-106.
36. Zhu Q, Oganov AR, Glass CW, Stokes HT. Constrained evolutionary algorithm for structure prediction of molecular crystals: methodology and applications. *Acta Crystallogr Sect B Struct Sci*. 2012;68:215-226.
37. Engh RA, Huber R. Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallogr Sect A*. 1991;47:392-400.

38. Rackers JA, Wang Z, Lu C, et al. Tinker 8: software tools for molecular design. *J Chem Theory Comput*. 2018;14:5273-5289.
39. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins Struct Funct Genet*. 2006;65:712-725.
40. Wang J, Cieplak P, Kollman PA. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J Comput Chem*. 2000; 21:1049-1074.
41. MacKerell AD, Bashford D, Bellot M et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B*. 1998;102:3586-3616.
42. Kaminsky GA, Friesner RA, Tirado-Rives J, Jorgensen WL. Evaluation and Reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J Phys Chem B*. 2001;105:6474-6487.
43. Lazaridis T, Karplus M. Effective energy function for proteins in solution. *Proteins Struct Funct Genet*. 1999;35:133-152.
44. Genheden S, Ryde U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin Drug Discovery*. 2015; 10:449-461.
45. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. *Proteins Struct Funct Genet*. 1995;23:566-579.
46. Valle M, Oganov AR. Crystal fingerprint space—a novel paradigm for studying crystal-structure sets. *Acta Crystallogr A*. 2010;66:507-517.
47. Bushlanov PV, Blatov VA, Oganov AR. Topology-based crystal structure generator. *CompPhys Comm*. 2019;236:1-7.
48. DeLano WL. *The PyMOL Molecular Graphics System, Version 1.8*. Schrödinger LLC; 2014.
49. AlQuraishi, M. AlphaFold2@CASP14: "It feels like one's child has left home." [Blog post]. (2020).
50. Finkelstein AV. Some additional remarks to the solution of the protein folding puzzle: reply to comments on "there and back again: two views on the protein folding puzzle". *Phys Life Rev*. 2017;21: 77-79.
51. Piana S, Klepeis JL, Shaw DE. Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. *Curr Opin Struct Biol*. 2014;24:98-105.
52. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*. 1970;57:97-109.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Rachitskii P, Kruglov I, Finkelstein AV, Oganov AR. Protein structure prediction using the evolutionary algorithm USPEX. *Proteins*. 2023;1-11. doi:10.1002/prot.26478