Research article

A protein folding potential that places the native states of a large number of proteins near a local minimum

Mukesh Chhajer[1] and Gordon M Crippen[*2]

Address: [1]Department of Chemistry, University of North Carolina, Chapel Hill, NC 27599, U.S.A and [2]College of Pharmacy, University of Michigan, Ann Arbor, MI 48109-1065, U.S.A

E-mail: Mukesh Chhajer - mchhajer@email.unc.edu; Gordon M Crippen* - gcrippen@umich.edu

*Corresponding author

Abstract

Background: We present a simple method to train a potential function for the protein folding problem which, even though trained using a small number of proteins, is able to place a significantly large number of native conformations near a local minimum. The training relies on generating decoys by energy minimization of the native conformations using the current potential and using a physically meaningful objective function (derivative of energy with respect to torsion angles at the native conformation) during the quadratic programming to place the native conformation near a local minimum.

Results: We also compare the performance of three different types of energy functions and find that while the pairwise energy function is trainable, a solvation energy function by itself is untrainable if decoys are generated by minimizing the current potential starting at the native conformation. The best results are obtained when a pairwise interaction energy function is used with solvation energy function.

Conclusions: We are able to train a potential function using six proteins which places a total of 42 native conformations within ~4 Å rmsd and 71 native conformations within ~6 Å rmsd of a local minimum out of a total of 91 proteins. Furthermore, the threading test using the same 91 proteins ranks 89 native conformations to be first and the other two as second.

Background

For the development of an all-encompassing potential energy function for the protein folding problem that can simulate both the thermodynamic and kinetic processes, it is argued that one must start from first principles [1]. However, given the complexity of the problem, involving thousands of atoms and an extremely large number of conformations available to these atoms, the available computational power posses a serious restriction. Alternatively, one could use the available experimental results to develop empirical potential functions using a simplified representation for the system which does not require an explicit enumeration of the entire system. Here, we present one such simplified method to train a potential energy function using a small set of proteins which places

a significant number of additional native conformations near its minima.

Ideally, we would like to obtain a potential energy function which assigns correct free energy to the native and non-native states for all the proteins. However, this requires knowing the complete distribution of energy states and total number of degeneracies for both the native and denatured states, quantities which are not readily available even for a simplified representation of our system. This has led to design of potential functions with restricted objectives.

Knowledge-based potentials are obtained from a survey of protein crystal structures. In this case, there is no training of parameters. The residue-residue interaction can be defined as contact vs. no-contact type[2–4] or, alternatively, can also be defined as distance dependent[5–7]. The resulting potentials depend upon the training set of proteins, reference state, and the functional form[8]. Thomas and Dill[9,10] have pointed out that these potential functions do not assign the correct value to the energy parameters, though they do rank them correctly.

On the other hand, the potential parameters are explicitly trained in the optimization-based methods. In these cases, the goal is to obtain a potential function which either maximizes the difference between the native and average non-native state energy (Z-score[11]), maximizes the ratio of folding to glass transition temperatures [12] or requires that all the non-natives be of energy higher than the native [13]. Here, the native state is represented by a single conformation. While the first two approaches provide reasonable results, they do not guarantee complete success in terms of fold recognition even for the training set proteins. A somewhat better criterion is to require that all the non-native conformations have energy higher than the native. In this case, if the energy function is chosen such that the total energy of the conformation is a linear function of its adjustable parameters, then the problem can be stated as a linear programming problem with a suitable objective function[13–20].

There are five factors that govern the training of a potential function using optimization-based methods: (i) representation of amino acids, (ii) choice of the interaction energy function, (iii) training set of proteins, (iv) generation of the alternative conformations or decoys and (v) the objective function.

The representation of the amino acids depends upon the amount of atomic-level detail one wishes to incorporate in the calculations, starting from a single point per residue representation[9–11] to an all-heavy atom representation[20,21]. As more and more details are included, the

complexity of the calculations increases; however, a coarse-grained model can not be expected to produce very refined structure predictions. While the lattice models have usually stayed with single point representation, the continuous state models have used different variations from a single point per residue to all-atom representation.

The choice of the interaction energy function to some extent depends upon the physical property of the amino acid which is deemed important, as well as on the type of amino acid representation used. The most commonly used pairwise inter-residue interactions can be treated as contact/no contact [14,15,22], discrete distance ranges [16,17,23], and continuously varying functions of distance [13,18,19,24,25]. The energy function should be flexible enough without causing overfitting. While having more parameters does provide the flexibility, having too many of them is not always helpful, as seen by Park et. al. [25] where a 80,000 parameter potential performed worse than a simple contact type function. Such discontinuous functions could also lead to problems if used for kinetic calculations or for local optimization. Furthermore, it has been shown that distance dependent energy functions perform better than contact type in a continuous conformation space [24].

Furthermore, excluded-volume effects play an important role in the performance of a potential function. As Thomas & Dill [10] have pointed out, one of the reasons for failure of the potential functions is the absence of any excluded-volume effects in the interaction energy models. Similarly, Park & Levitt [24] show that a van der Waals type energy function gives better results than contact type or solvation potential alone, and a combination of surface energy function with a van der Waals type energy function performs even better.

Another crucial requirement for the development of potential functions using optimization-based methods is the generation of realistic and challenging decoys. A great deal depends upon the set of decoys used to define the non-native state. It has been demonstrated by various investigators [14,19,26] that using a set of decoys obtained mostly by threading does not enforce a very stringent training criterion, and the potential functions so obtained fail to even place the training set native conformations near a local minimum, let alone the global minimum. Similarly, one can obtain low energy conformations by manipulation of conformations in a contact map representation, though this may lead to geometrically unrealizable conformations [14]. Other methods which have been used to generate low energy decoys include Monte-Carlo simulations [27] and molecular dynamics simulations [28,29], both of which require significant computation time, discretiza-

tion of conformational variables [24], and inclusion of rigid pieces of structures [30].

A necessary, though not sufficient, condition for a well behaved potential function is that the native conformation be at or near a minimum of the potential function. As pointed out earlier [14,19,26], the potential functions derived with mostly threaded conformations fail to satisfy this condition, even when a large set of conformations is used. The reason for such a failure is that the training conformations are fixed and not allowed to relax given a potential function. This results in most of the conformations being of significantly higher energy and only a small set of conformations end up providing an active constraint set. However, if the inactive conformations are allowed to relax, they would provide a much more stringent set of constraints and will improve the performance of the potential function significantly. In this work, we place a milder condition on the training by only asking that the native be near a local minimum. No condition is placed for the global minimum. We use energy minimization to generate new decoys which are physically realizable low energy decoys and provide better training. Furthermore, a physically meaningful objective function is used during the quadratic programming. Here we train a potential function using only a few single domain, monomeric proteins which do not require any hetero groups or ligands to stabilize their folded state. These chains have less than 10% sequence alignment (using MOE [31]) and significantly different crystal structures (average rmsd of ~11 Å). Once trained, the potential function is able to place 42 proteins, even including some multi-chain proteins and two CASP3 proteins, to within ~4 Å of the experimental native conformations. Here we present a method that trains a reasonably good potential function using only a few proteins.

Results and Discussion

As we have shown in an earlier work [26], the generation of good decoys is crucial for the training of the potential function. The potential functions trained mostly by the threaded conformations, even using a large set of training proteins, do not put the native conformations of even the training set proteins at a local minimum or near it, a necessary, though not a sufficient, condition for the stability of the native conformation. This is a serious shortcoming of the potential function, since without ensuring the local minimum condition, there is no hope of ensuring that the native is the global minimum conformation, much less that the native is thermodynamically stable. Here we use energy minimization starting from the native conformation to generate new decoys which depend upon the current potential function. However, since this process is much more time consuming than using a library of conformations, we use a small set of proteins to train our potential function, though the training is much more

rigorous. We also use small proteins since computation time increases roughly as a square of the chain length. The proteins we use have very little sequence identity and have very dissimilar crystal structures. This allows for various residues to be in different environments. For the first set of three proteins (Set A: {1ayi, 1bk2, 1ubi}; see Table 1), we perform calculations using all the three energy functions to find the one which works best and then use that energy function to get a better potential function using a slightly expanded set of proteins. All the proteins we use in the training set are single chain, single domain proteins which do not require any hetero groups or any other ligands to form a stable folded state in the aqueous solution. The potential function so trained is then used on the test set to evaluate its performance. The test set contains some close homologs but mostly quite distinct proteins compared to the training set in terms of their size, sequence, and crystal structures.

Table 1: Effect of Solvation with three training proteins

protein	without solvation (PIE3)		with solvation (PSE3)	
	ρ	rmsd (Å)	ρ	rmsd (Å)
1ayi	0.165	2.03	0.150	1.87
1bk2	0.212	2.04	0.118	1.13
1ubi	0.128	1.47	0.168	1.94
1ame	0.386	3.99	0.606	6.04
1ayd	0.492	6.39	0.440	5.60
1bm8	0.651	8.17	0.390	4.79
1enh	0.363	3.55	0.324	3.27
1gzi	0.355	3.66	0.456	4.43
1igd	0.928	11.15	0.220	2.50
1mjc	0.750	8.45	0.313	3.29
1pgb	0.450	4.73	0.278	2.82
1ptf	0.820	9.85	0.450	5.07
1pwt	0.296	3.03	0.187	1.90
1szt	0.867	11.50	0.181	2.92
1vcc	0.625	7.67	0.423	4.91

There are two additional constants, G and ρ_{cut} (described in Parameter Adjustment section), which need to be fixed before the calculation can be performed. In our previous calculations [26], we had fixed $G = 0.3$ and $\rho_{cut} = 0.10$. The value of G defines the minimum energy separation between the native and an alternate conformation. We would like this value to be large so that the native is stable. The value of G depends upon the kind of energy function being used and in this work, we vary the value between 0.3 and 60.0. Similarly, the value of ρ_{cut} defines the basin of the native conformation and this value is varied be-

tween 0.05 and 0.25. To obtain the best potential function, i.e., a potential function that places the maximum number of native conformations of test set proteins close to its minima, we vary these two parameters and repeat the training process.

PIE function
The PIE function, shown in Figure 1, has 3 parameters per interaction type, hence a total of 900 parameters. The potential function was so trained that when the native conformations of the training set proteins are energy minimized, the resulting local minimum conformation is within 2.5 Å rmsd. For all choices of $0.3 \leq G \leq 60.0$ and $0.05 \leq \rho_{cut} \leq 0.25$, the potential was trainable, i.e., the potential function had enough flexibility that we did not run in to a situation where the QP could not come up with a solution, though at times solutions were suboptimal. The performance of these potential functions on the test set proteins varied greatly. In general high G and low ρ_{cut} values resulted in a very rough energy surfaces which led to large energy changes for small conformational variations. On the other hand, low G and high ρ_{cut} values resulted in a potential function that is too flat and allows easy conformational changes. For these proteins, we found that a value of G = 10.0 and ρ_{cut} = 0.12 gave us the best results. This potential function (PIE3) is able to place four of the 12 test protein native conformations to within ~4 Å rmsd of a local minimum (see Table 1). The other eight are between 4–12 Å rmsd.

SE function
The SE function is very simple and has only 21 adjustable parameters, one for each residue and one for the overcrowding which prevents a complete collapse of the chain as well as prevents any unrealistic chain overlaps. Once again, we used the same three proteins and used various combinations of G and ρ_{cut} to obtain solutions. However, this time, with the SE function by itself, we failed to obtain a potential function even for a single protein, let alone all three proteins together. As soon as the conformations were generated by energy minimization process, we obtained a set of inequalities which resulted in a null feasible solution region for QP. This shows that the SE function by itself does not have enough flexibility to give a trained potential function when decoys are generated by more rigorous methods. Yet, this function was successfully trained by threading using experimental crystal structures [18].

PSE function
As with the PIE, the same three proteins are used for the training purposes. Since this is a sum of the other two energy functions, there are 920 parameters, the one parameter for overcrowding in the SE function having been removed since the f_0 part of the PIE function performs a similar function. Once again, we obtain various sets of
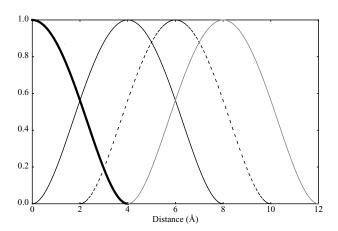

Figure 1
Terms of pairwise interaction energy function, eq. (2). Heavy solid line is f_0, thin solid line is f_1, dashed is f_2, and dotted is f_3.

trained parameters depending upon the value of G and ρ_{cut}, and once again we find that G = 10.0 and ρ_{cut} = 0.12 give the best results for the test proteins. In this case (PSE3), six of these proteins remain within ~4 Å rmsd and all 12 are within ~6 Å. Considering that the test set contains both close homologs and significantly different proteins from the training proteins in terms of their size, sequence, and structure, this is a very encouraging result since all 12 test proteins remain within ~6 Å compared to ~12 Å for PIE. It shows that using a PSE function, one can train a potential function using only a few proteins which would be applicable to many other proteins. The comparison between the results of the PIE and PSE3 functions is shown in Table 1.

The training for PSE3 was performed using three proteins (Set A = {1ayi, 1bk2, 1ubi}), none of which contains Cys. However, there were many other interactions which remained untrained in the final version of PSE3. To get all 300 pairwise interactions to train at least moderately, we had to either select larger proteins which contain all residues in sufficient number or enlarge the training set even if all the residues are not present in any one of the protein. Since computation time increases as square of the chain length, we chose to enlarge the training set using smaller proteins. A set of four proteins (Set B: {1enh, 1bk2, 2era, 1ubi}) contains all the 20 amino acids collectively. However, in the final trained potential function (PSE4), eight of the interactions still remain completely untrained. This set was then further expanded by including two additional proteins (Set C: {1enh, 1bk2, 2era, 1ubi, 1ail, 1dsl}). The final trained potential function using Set C (PSE6) has all the pairwise interactions at least mildly trained. Set C contains two proteins (1enh and 1ail) with mostly α-helical

structure, three (1bk2, 2era and 1dsl) with mostly β-sheet structure and one (1ubi) with α/β structure. However, there is very little sequence identity between any of these proteins, and their crystal structures are also very different (less than 10% sequence alignment and rmsd of 8–14 Å in their crystal structures using MOE). The corresponding solvation parameters for various residues are shown in Table 2.

Table 2: Solvation parameters

Residue	PSE4	PSE6
Gly	0.39958884300	-0.17727756470
Ala	0.86811242830	-0.36213127280
Val	0.62981201440	-0.36741883040
Leu	0.36371810330	0.82156988920
He	0.89053343600	0.09748013457
Cys	1.39536632100	0.09762673560
Met	-0.42954544170	0.15720350550
Phe	-0.01532194602	-0.22049718860
Pro	1.60142719700	0.03948009827
Tyr	-0.37437233800	0.71270611040
His	0.45098491050	-0.44487856400
Trp	1.01909618300	0.26052565020
Ser	-0.19148377760	-0.44395269350
Thr	0.47362962090	0.11467398110
Lys	-0.14119705180	-0.15725262960
Arg	-0.30176394520	-0.30074012620
Asp	0.42807283590	-0.58936068250
Asn	0.30127878790	0.52122075820
Glu	-0.19447595190	0.69933440120
Gln	-0.58651775250	0.01869831326

The best results for Set B (PSE4) are obtained using $G = 9.9$ and $\rho_{cut} = 0.12$. The potential function PSE4 places 14 of the 15 test set native conformations to within ~4 Å of a local minimum (see Table 3) which is a significant improvement over PSE3. The PSE4 was further applied to another 72 proteins (see Table 4) for a total of 87 test proteins. Of these 87 test proteins, 31 native conformations were placed within ~4 Å rmsd of a local minimum (including three CASP3 proteins), and 59 are within ~6 Å rmsd. Thus, PSE4 places a total of 35 native conformations (including the training set proteins) of very different sequences and crystal structures to within 4 Å rmsd and 63 to within 6 Å rmsd of a local minimum out of a total of 91 proteins. The set of 91 proteins contains a variety of proteins, some of which are multi-chain (e.g., 1d3b) or require a large group for stabilization (e.g., 1cc5).

The best result for Set C (potential function PSE6 with $G = 13.1$ and $\rho_{cut} = 0.15$) places 11 out of 14 test set proteins to within ~4 Å rmsd (see Table 3). In all, out of 91 pro-

teins including the training set, 71 are within ~6 Å and 42 are within ~4 Å (including two CASP3 proteins). The fact that PSE6 is able to place 71 out of 91 native conformations within ~6 Å rmsd of a local minimum for proteins of very different sequences, crystal structures, and length points to the robustness of the potential function, and, therefore, the usefulness of the approach.

Threading test
To further test the validity and usefulness of the final potential function PSE6, we performed an ungapped threading test using all the 91 proteins. We first energy minimized each of the native conformations and obtained the conformation of the local minimum using potential PSE6. These local minimum conformations were then used as decoys during ungapped threading tests for each protein. The threaded conformations were not energy minimized. In the threading test, instead of the usual single point representation for each residue, we use a five point representation for each residue, since the potential PSE6 has been trained using this representation. Out of 91 proteins, 89 native conformations are ranked first and the other two (1dsl, 1snc) are ranked second. This test is conducted using a total of 180,878 decoys. Therefore, we see that a potential function trained using the procedure outlined in this work also performs well when subjected to the threading test, whereas the potentials designed using only threaded conformations do not perform well when decoys are generated using energy minimization.

One of the drawbacks of this procedure is that the training process is not sequential and, therefore, quite time consuming, i.e., addition of inequalities for a new protein after the training for one protein is complete does not necessarily give better results. Any change in protein set, G, ρ_{cut} or any other strategy change requires starting all over again, since the set of conformations generated for the training are based on the current set of parameters, i.e., the process is memory dependent. However, given that the potential energy surface is of very high dimension and highly irregular, it is to be expected.

By only considering the energy minimization of the experimental native conformation, we reduce the amount of time required to train the potential function. Each potential function could be trained within a week running on four Sun workstations (CPU speeds: 135 MHz-450 MHz). However, we had to perform many training runs to get the best possible potential function for each set, and at this point it guarantees only the local minimum condition for the native state, not the global one.

Conclusions
A necessary, though not sufficient, condition for the stability of the native state of a protein is that the experimen-

Table 3: Comparison of PSE4 and PSE6

protein	fold type	chain length	PSE4		PSE6	
			rmsd (Å)	ρ	rmsd(Å)	ρ
1ail	α	70			2.24	0.178
1bk2	β	57	1.65	0.173	2.45	0.251
1dsl	β	87			2.10	0.181
1enh	α	54	1.63	0.160	2.19	0.216
1ubi	α/β	76	1.03	0.090	2.41	0.211
2era	β	62	1.22	0.106	2.36	0.212
1a19.A	α/β	90	3.03	0.256	3.20	0.264
1ame	α/β	67	4.02	0.400	3.76	0.375
1ayd	α/β	101	3.53	0.288	3.44	0.276
1ayi	α	87	3.96	0.323	5.42	0.438
1bm8	α/β	99	2.90	0.237	4.97	0.409
1dsl	β	87	3.98	0.362		
1gzi	α/β	65	3.21	0.327	3.31	0.334
1igd	α/β	61	2.66	0.234	3.33	0.290
1mjc	β	69	3.50	0.334	2.90	0.278
1ops	α/β	64	2.38	0.238	3.46	0.347
1pgb	α/β	56	3.01	0.296	2.88	0.275
1ptf	α/β	88	4.58	0.421	3.86	0.340
1pwt	β	61	2.17	0.211	3.39	0.323
1szt	α	68	2.99	0.187	7.91	0.533
1vcc	α/β	77	3.67	0.307	3.36	0.287
1yhb	β	87			7.44	0.523

tal native conformation be at or near a minimum of the potential function. We show that by using only a small set of proteins, potential functions can be trained that put a large number of native states near a local minimum of the potential function. While this does not guarantee the stability of the native state, this does ensure that the native conformation is not sitting at highly unstable points on the potential surface, a condition which is encountered in other potential functions. Our best potential function (PSE6) obtained using training Set C (containing six proteins) is able to place a total of 42 native conformations to within ~4 Å rmsd and 71 native conformations to within ~6 Å rmsd of a local minimum of the potential function out of the 91 proteins. We also find that while the pairwise interaction energy function is trainable using the more rigorously generated decoys, the solvation potential alone is not. The solvation potential has only 21 adjustable parameters and can not provide enough flexibility to satisfy all the constraints in our case, though it was trainable using threaded conformations from PDB. The best results are obtained using the pairwise energy function in combination with the solvation energy function. We further test the final potential function PSE6 with threaded conformations for 91 proteins and find that 89 of the native conformations are ranked first and the remaining two are ranked second. The training process is further improved by the use of a physically meaningful quantity for the objective function in the QP optimization which helps in identifying a better solution.

Methods
Generation of native structure
For the purposes of training, we select only single domain, monomeric proteins which do not require any hetero atoms or ligands for their stabilization in the folded state. Furthermore, the atom positions are available for the main chain heavy atoms and at least the corresponding C_β atom of the residue. The crystal structures (obtained from PDB[32]) for the protein chains are fitted to a standard geometry continuous state model (very similar to the one used by Dill *et. al.*[33]). In the fitted model, each side chain is represented by a single interacting site, located at the C_β, while keeping all the main chain heavy atoms, i.e., each residue is represented by five interacting sites (united atom types). Standard values for bond lengths and bond angles are used, and all peptide bonds are kept in the trans conformation. Thus, only torsion angles (ϕ, ψ) are allowed to vary between -180° and 180°. The details of the

Table 4: Comparison of PSE4 and PSE6 (additional tests)

protein	chain length	PSE4		PSE6	
		rmsd (Å)	ρ	rmsd (Å)	ρ
1acp	77	4.88	0.436	6.22	0.577
1ail	70	5.91	0.475		
1amm	174	6.86	0.464	3.72	0.238
1bdo	79	3.29	0.289	3.30	0.288
1beo	98	4.83	0.381	3.77	0.304
1ble	161	4.85	0.338	4.16	0.280
1bor	56	6.95	0.750	5.42	0.619
1c94.A	37	2.48	0.149	5.10	0.318
1cc5	83	7.04	0.657	13.89	1.273
1cd8	114	4.18	0.321	4.88	0.363
1cfe	135	6.50	0.474	7.87	0.543
1chd	198	4.11	0.279	4.96	0.333
1cm	46	3.91	0.396	4.30	0.484
1tf	68	2.52	0.248	3.20	0.302
1cyo	88	3.83	0.320	5.94	0.492
1d3b.A	72	2.93	0.268	2.63	0.240
1d3b.B	81	3.65	0.275	3.60	0.261
1div	149	11.11	0.473	11.06	0.489
1ecd	136	5.38	0.375	5.02	0.367
1exg	110	3.61	0.274	4.03	0.303
1f3g	150	4.88	0.357	3.83	0.279
1fbr	93	9.01	0.604	6.61	0.428
1fdx	54	9.29	0.869	5.72	0.703
1fkf	107	5.59	0.450	3.56	0.279
1fxd	58	4.01	0.428	3.23	0.341
1hfh	120	9.37	0.527	4.12	0.227
1hoe	74	3.12	0.285	6.47	0.560
1ife	131	4.53	0.341	3.91	0.278
1jpc	108	5.55	0.399	6.01	0.427
1knb	186	5.18	0.320	4.64	0.291
1kuh	132	8.71	0.653	6.68	0.477
1lba	146	8.15	0.598	3.90	0.288
1lzl	130	5.18	0.388	4.33	0.324
1mai	119	4.99	0.380	3.73	0.280
1paz	120	3.87	0.305	3.94	0.317
1pcy	99	6.70	0.537	5.29	0.447
1pdo	129	4.66	0.342	10.08	0.691
1pkp	145	4.19	0.287	7.10	0.486
1poa	118	8.65	0.681	6.70	0.517
1poc	134	13.60	0.982	8.66	0.587
1r69	63	5.19	0.517	3.21	0.311
1ra9	159	7.24	0.497	3.34	0.220
1rie	127	5.17	0.393	4.98	0.364
1rsy	135	6.47	0.439	4.98	0.328
1skz	104	6.60	0.458	7.04	0.512
1snc	135	3.76	0.275	6.96	0.492
1vhh	157	5.12	0.361	6.81	0.465
1whi	122	3.49	0.264	3.75	0.280
1xnb	185	3.22	0.213	3.47	0.229
1ycc	108	6.50	0.515	5.35	0.429
1yua	122	8.53	0.626	7.49	0.489
2abd	86	8.38	0.752	5.21	0.432
2 end	137	8.20	0.534	5.71	0.356
2fx2	147	4.18	0.309	4.87	0.350
2fxb	81	4.08	0.382	3.40	0.322

Table 4: Comparison of PSE4 and PSE6 (additional tests) *(Continued)*

2hbg	147	5.72	0.400	3.99	0.278
2lhb	29	9.24	0.703	3.42	0.299
2mcm	112	3.77	0.303	4.87	0.360
2rhe	114	3.47	0.263	4.25	0.315
2m2	155	7.15	0.507	8.33	0.542
2sns	141	6.61	0.459	7.74	0.522
2stv	184	6.31	0.366	6.91	0.421
2tgi	112	7.88	0.463	4.05	0.238
3b5c	85	5.62	0.471	5.23	0.443
3chy	128	3.30	0.261	3.46	0.268
451c	82	5.65	0.487	5.74	0.514
4fd1	106	6.15	0.510	3.27	0.286
4fxn	138	4.78	0.357	5.43	0.415
4icb	76	4.67	0.431	3.85	0.340
5rxn	54	4.93	0.509	4.10	0.421
7rsa	124	8.70	0.605	5.74	0.404

fitting procedure are described in ref. 20 and will not be repeated here. The fitted model is within 0.5 Å rmsd [34] of the PDB structure and is used as the native structure in our calculations. Thus, the native state is a single energy, non-degenerate state in our calculations.

Generation of the non-native decoys
The conformations of the non-native decoys are generated using the following procedures. (i) Parameter-Independent Decoys: Starting from the native conformation, all pairs of torsional angles are perturbed by $-30° \leq \delta \leq 30°$ at a time. For example, for a chain of 50 residues, there are 100 torsion angles and, therefore, 4950 pairs of torsional angles. This would give us 4950 different decoys, but see the Solution Procedure section for the precise protocol used. These conformations do not depend upon the current potential function. (ii) Parameter-Dependent Decoys: The native conformation is energy minimized with respect to torsion angles using the current potential function.

Energy functions
Since we generate conformations by energy minimization with respect to the torsion angles which are allowed to change continuously, we also need an energy function that is a continuous and differentiable function of angles. We use three different types of energy functions: the pairwise interaction energy (PIE) function, the solvation energy (SE) function, and the pairwise interaction with solvation energy (PSE) function. In our case, all these functions are continuous functions of separation between atoms and, therefore, the conformations can be energy minimized with respect to torsion angles to generate decoys once we use the standard geometric representation for the conformations.

Pairwise interaction energy (PIE) function
The pairwise interaction between a pair of atoms is represented by [19]

$$e_{t_i t_j}\left(x_{ij}\right) = \sum_{k=0}^{3} p_{k, t_i t_j} f_k\left(x_{ij}\right) \quad (1)$$

where $p_{k, t_i t_j}$ are adjustable parameters, t_i is the type of united atom i, and x_{ij} is the distance between united atoms i and j. The function $f_k(x_{ij})$ is defined over a limited distance interval as follows:

$$f_k\left(x_{ij}\right) = \begin{cases} \left(\left(x_{ij} - a_k\right)^2 - b^2\right)^2 / b^4 & for\ a_k - b \leq x_{ij} \leq a_k + b \\ 0 & otherwise \end{cases} \quad (2)$$

where $b = 4.0$, $a_0 = 0.0$, $a_1 = 4.0$, $a_2 = 6.0$ and $a_3 = 8.0$. The first term in each interaction, $f_0(x_{ij})$, provides some steric repulsion for atoms sitting too close to each other and its coefficient is always fixed at 10 which prevents a complete collapse of the chain. The other three functions have adjustable coefficients which are allowed to vary between -20 and 20. The functional form of $f_k(x_{ij})$'s is shown in Figure 1. Each interaction energy is represented by four parameters. The maximum interaction distance is 12 Å. Equation (1) has following properties:

$$e_{t_i t_j}\left(0\right) = 10$$

$$e_{t_i t_j}\left(x_{ij}\right) = 0\ for\ x_{ij} \geq 12\,\text{Å}$$

$$\frac{\partial}{\partial x_{ij}} e_{t_i t_j}\left(x_{ij}\right)|_{x_{ij}=12} = 0$$

The interaction energy function $e_{t_i t_j}\left(x_{ij}\right)$ is continuous and differentiable for all $x_{ij} \geq 0$, and is a linear function of adjustable parameters.

Furthermore, we consider 24 atom types: 20 side chain atom types representing 20 different amino acids and four main chain heavy atoms (C_α, C, N, and 0). This gives us a total of $24 \times 25/2 = 300$ different types of interactions and a total of 1200 parameters. In the present study, 300 of these parameters have a fixed value of 10, and 900 are adjustable. The total pairwise energy is

$$E_{PIE} = \frac{1}{2} \sum_{|i-j|>2} e_{t_i t_j}\left(x_{ij}\right) \tag{3}$$

where the summation is over all pairs of united atom types separated by at least three covalent bonds. Since the total pairwise energy of any conformation is just the sum of individual pairwise interaction energies, the total pairwise energy is also a linear function of the adjustable parameters.

Solvation energy (SE) function
The solvation energy represents the change in energy due to the burial of various residues. This is an average contribution for each residue based upon how many other residues are surrounding it and given by

$$E_{SE} = \sum_{|i-j|>2} p_{k(i)} C_{t_i t_j}\left(x_{ij}\right) + \sum_i P_S S_i \tag{4}$$

where $p_{k(i)}$ is the energy contribution for a residue of type k at sequence position i along the chain surrounded by other residues, and P_s is the penalty for overcrowding. Here, the first summation is over all residue pairs separated by at least two residues along the chain backbone, and second summation is over all those residues for which the total number of contacts exceeds the maximum allowable number. The number of excess contacts S_i are

$$S_i = max\left[0, \sum_j C_{t_i t_j}\left(x_{ij}\right) - C_{k(i),\,max}\right] \tag{5}$$

The function $C_{t_i t_j}\left(x_{ij}\right)$ determines the burial of each residue where x_{ij} is the separation between two side chain C_β atoms situated at positions i and j. These are sigmoidal functions (see Figure 2). Both $C_{t_i t_j}\left(x_{ij}\right)$ and $C_{k(i),\,max}$ were obtained by Dombkowski and Crippen[18] from the survey of PDB crystal structures, and these functions are predetermined in our training procedure. Only the $p_{k(i)}$'s and P_S are determined here. This energy function has therefore only 21 adjustable parameters.
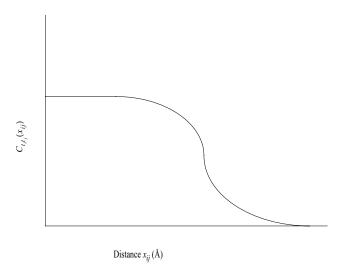

Figure 2
Functional form of $C_{t_i t_j}\left(x_{ij}\right)$, eq. (4).

Pairwise interaction with solvation energy (PSE) function
The PSE function includes both the pairwise interaction energy and the solvation energy without the overcrowding part since the $f_0(x_{ij})$ part of the PIE function ensures that the chain does not collapse to a point. The total energy in this case is given by

$$E_{PSE} = E_{PIE} + \sum_{|i-j|>2} p_{k(i)} C_{t_i t_j}\left(x_{ij}\right) \tag{6}$$

and it has a total of 920 adjustable parameters.

Parameter adjustment using quadratic programming
Given the above energy functions, the total energy of any conformation is a linear function of the adjustable parameters, irrespective of the energy function being used. This property allows us to use quadratic programming (QP) to adjust the values of these parameters. We choose the following objective function $\mathcal{F}$ for the optimization purposes:

$$\mathcal{F} = \sum_{k=1}^{P} \left(\nabla E\right)_{k,\,nat}^2 + \sum_{k=1}^{N_P} p_k^2 = \sum_{n=1}^{P}\left(\sum_{k=1}^{N_P} \alpha_{n,k} p_k\right)^2 + \sum_{k=1}^{N_P} p_k^2 \tag{7}$$

where $\alpha_{n,k}$ is the sum of the coefficients of the N_p parameters p_k when the derivative of the energy of a conformation is evaluated at the native conformation of the n-th protein. While the first term minimizes the gradient of the energy with respect to dihedral angles at the native conformation for all training set proteins, the second term keeps the values of parameters small to reduce the roughness of the energy surface. Having a continuous energy function is essential for using such a physically meaningful objective function. By minimizing the magnitude of the gradi-

ent of energy of the native conformation, we obtain a set of parameters such that the native conformation approximates a stationary point of the potential function. The total number of parameters N_P depends upon the type of energy function is being used, namely 900 for the PIE function, 21 for the SE function, and 920 for the PSE function. The constraints are of the form

$$\Delta E = E_{non} - E_{nat} > g \qquad (8)$$

where

$$g = \begin{cases} G & \rho > 0.3 \\ G\dfrac{\rho}{0.3} & \rho_{cut} < \rho \le 0.3 \end{cases} \qquad (9)$$

where G and ρ_{cut} are adjustable constants. If $\rho < \rho_{cut}$, we consider the decoy to be conformationally identical to the native and no inequality is generated even if the energy of the decoy is lower than that of the native. The constants G and ρ_{cut} define the minimum energy separation between the native and decoys, and the radius of the basin of the native, respectively. They may be changed from one training run to another; however, not during the same training run. Here, ρ is given by [35]

$$\rho = \frac{2 \cdot D_{1,2}}{\sqrt{2\left(R_{g1}^2 + R_{g2}^2\right) - D_{1,2}^2}} \qquad (10)$$

where R_{gj} is the radius of gyration of the jth conformation and $D_{1,2}$ is the customary root mean square deviation in C_α coordinates after optimal superposition of the two conformations, taken to be the native and an alternate. By allowing g to vary linearly with ρ near the native structure, we avoid sudden jumps in the potential energy surface and keep the variation in the energy function smooth near the native. Finally, we need to provide bounds on our parameters, and we have chosen $p_k = 10$ for all p_k's which multiply $f_0(x_{ij})$'s and $-20 \ge p_k \le 20$ for the rest of them. These values are chosen so that we are able to find a feasible solution while keeping most of the parameters away from the extreme values. Due to the choice of $\mathcal{F}$, we use quadratic programming (QP) instead of linear programming to solve the set of inequalities. [36]

Solution procedure

We obtain the initial set of parameters by minimizing $\mathcal{F}$ using QP. The decoys are generated using the methods described earlier. First, the decoys are generated by perturbing pairs of torsion angles by an amount δ. For a chain of n residues, there are $2n(2n\text{-}1)/2$ such pairs. Even for a chain with 50 residues, this would produce 4950 conformations. Initially, since the parameter set is completely untrained, a large number of inequalities are generated for a training set of three to six proteins ranging in length from 50–90 residues. To keep the number of inequalities to a reasonable value, we first perturb pairs of angles sep-

arated by a certain number of covalent bonds n and its multiples. For example, in the first pass, we change the pair of angles separated by $n = 9$ covalent bonds or multiples of it, i.e., 9, 18, 27, etc. In this case, starting from the fifth torsion angle from the N-terminus, we perturb pairs of angles located at $(5+9k, 5+9l)$ positions along the chain where $k \le l = 0$, 1, 2, *etc*, collect all the inequalities and run them through the QP to obtain a new set of parameters. Next, we take pairs of angles separated by 4 backbone bonds and add new inequalities to the previous set and obtain a new set of parameters by QP. This process is repeated for pairs of angles separated by two, one and zero backbone bonds. Next, we repeat this process for all pairs of angles by perturbing them by $\delta \pm 2°$, $\delta \pm 4°$ and $\delta \pm 6°$ where $12° \le |\delta| \le 20°$. This is the parameter independent generation of decoys.

The next step is parameter dependent generation of decoys. Here, we use energy minimization of native conformation (using *BFGS* method [37]) to generate new decoys. During each cycle, we begin from the native conformation and energy minimize the conformations with respect to the torsion angles. After every 100 minimization steps, the conformational distance of the current conformation from the native is calculated in terms of ρ. If the ρ value is greater than the cutoff value ρ_{cut}, we say that the current conformation is sufficiently different from the native one, i.e., it does not belong to the basin of native, and since the energy of this conformation is lower than the native, a new constraint is generated. After complete minimization of the native conformations, if any new inequalities are generated, they are added to the set of inequalities and a new set of parameters is obtained by QP. If no new inequalities are generated, then the training is complete and a trained potential has been obtained. To keep the number of inequalities within a manageable limit, we periodically discard inequalities corresponding to nonnatives having energy significantly higher than the native state, i.e., having high slack value greater than S for the current parameter set. Initially, S is set at 500 and slowly reduced to 200.

Authors' Contributions

M.C. designed the fitting protocols, wrote the computer programs, chose training and test data, and carried out the calculations. G.M.C. contributed overall planning, advice, and discussions throughout. Both have read and approved the manuscript.

Acknowledgements

References

1. Lazaridis T, Karplus M: **Effective energy functions for protein structure prediction.** *Curr Opin Struct Biol* 2000, **10:**139-145
2. Miyazawa S, Jernigan R: **Estimation of effective contact energies from protein crystal structures: Quasi-chemical approximation.** *Macromolecules* 1985, **18:**534-552
3. Miyazawa S, Jemigan R: **Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading.** *J Mol Biol* 1996, **256:**623-644
4. Hinds D, Levitt M: **Exploring conformational space with a simple lattice model for protein structure.** *J Mol Biol* 1994, **243:**668-682
5. Skolnick J, Jaroszewski L, Kolinski A, Godzik A: **Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct?** *Protein Sci* 1997, **6:**676-688
6. Sippl M: **Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures.** *J Comput-Aided Mol Design* 1993, **7:**473-501
7. Kocher J-P, Rooman M, Wodak S: **Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches.** *J Mol Biol* 1994, **235:**1598-1613
8. Godzik A, Kolnski A, Skolnick J: **Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets.** *Protein Sci* 1995, **4:**2107-2117
9. Thomas P, Dill K: **An interactive method for extracting energy-like quantities from protein structure.** *Proc Natl Acad Sci USA* 1993, **93:**11628-11633
10. Thomas P, Dill K: **Statistical potentials extracted from protein structures: How accurate are they?** *J Mol Biol* 1996, **257:**457-469
11. Mirny L, Shakhnovich E: **How to derive a protein folding potential? A new approach to an old problem.** *J Mol Biol* 1996, **264:**1164-1179
12. Goldstein R, Luthey-Shulten Z, Wolynes P: **Protein tertiary structure recognition using optimized Hamiltonians with local interactions.** *Proc Natl Acad Sci USA* 1992, **89:**9029
13. Maiorov V, Crippen G: **Contact potential that recognizes the correct folding of globular proteins.** *J Mol Biol* 1992, **227:**876-888
14. Vendruscolo M, Domany E: **Pairwise contact potentials are unsuitable for protein folding.** *J Chem Phys* 1998, **109:**11101-11108
15. Vendruscolo M, Najmanovich R, Domany E: **Can a pairwise contact potential stabilize native protein folds against decoys obtained by threading?** *Proteins* 2000, **38:**134-148
16. Tobi D, Elber R: **Distance-dependent, pair potential for protein folding: Results from linear optimization.** *Proteins* 2000, **41:**40-46
17. Tobi D, Shafran G, Linial N, Elber R: **On the design and analysis of protein folding potentials.** *Proteins* 2000, **40:**71-85
18. Dombkowski A, Crippen G: **Disulfide recognition in an optimized threading potential.** *Protein Eng.* 2000, **13:**679-689
19. Ohkubo Y, Crippen G: **Potential energy function for continuous state model of globular proteins.** *J Comput Biol* 2000, **7:**363-379
20. Crippen G: **Constructing smooth potential functions for protein folding.** *J Mol Graph Mod* 2001, **19:**87-93
21. Samudrala R, Moult J: **An all-atom distance dependent conditional probability discriminatory function for protein structure prediction.** *J Mol Biol* 1998, **275:**895-916
22. Clementi C, Vendruscolo M, Maritan A, Domany E: **Folding Lennard-Jones proteins by a contact potential.** *Proteins* 1999, **37:**544-553
23. Park K, Vendruscolo M, Domany E: **Towards an energy function for the contact map representation of proteins.** *Proteins* 2000, **40:**237-248
24. Park B, Levitt M: **Energy functions that discriminate X-ray and near native folds from well-constructed decoys.** *J Mol Biol* 1996, **258:**367-392
25. Park B, Huang E, Levitt M: **Factors affecting the ability of energy functions to discriminate correct from incorrect folds.** *J Mol Biol* 1997, **266:**831-846
26. Chhajer M, Crippen G: **A protein folding potential that has a stable native state and correct free energy of unfolding,** 2002
27. Micheletti C, Seno F, Banavar J, Maritan A: **Learning effective amino acid interactions through interactive stochastic techniques.** *Proteins* 2001, **42:**422-431
28. Wang Y, Zhang H, Li W, Scott R: **Discriminating compact non-native structures from the native structure of globular proteins.** *Proc Natl Acad Sci USA* 1995, **92:**709-13
29. Huang E, Subbiah S, Tsai J, Levitt M: **Using a hydrophobic contact potential to evaluate native and near-native folds generated by molecular dynamics simulations.** *J Mol Biol* 1996, **257:**716-725
30. Crippen G, Ohkubo Y: **Statistical mechanics of protein folding by exhaustive enumeration.** *Proteins* 1998, **32:**425-437
31. **MOE, Version 1999.09.** *Chemical Computing Group Inc* 1999 [http://www.chem-comp.com]
32. Berman H, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov I, Bourne P: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28:**235-242
33. Dill K, Phillips A, Rosen J: **Protein structure and energy landscape dependence on sequence using a continuous energy function.** *J Comput Biol* 1997, **4:**227-239
34. Kabsch W: **A discussion of the solution of the best rotation to relate two sets of vectors.** *Acta Cryst* 1978, **A34:**827-828
35. Maiorov V, Crippen G: **Size-independent comparison of protein 3-dimensional structures.** *Proteins* 1995, **22:**273-283
36. **CPLEX, Version 6.6. ILOG, Inc** 2000 [http://www.ilog.com]
37. Shanno D, Phua K: **Remark on "Algorithm 500: Minimization of unconstrained multi-variate function [E4]".** *ACM Trans Math Software* 1980, **6:**618-622