Structural Bioinformatics

KORP: Knowledge-based 6D potential for fast protein and loop modeling

José Ramón López-Blanco[1*] and Pablo Chacón[1*]

[1]Department of Biological Chemical Physics, Rocasolano Institute of Physical Chemistry C.S.I.C., Serrano 119, 28006 Madrid, Spain

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Knowledge-based statistical potentials constitute a simpler and easier alternative to physics-based potentials in many applications, including folding, docking, and protein modeling. Here, to improve the effectiveness of the current approximations, we attempt to capture the 6-dimensional (6D) nature of residue-residue interactions from known protein structures using a simple backbone-based representation.

Results: We have developed KORP, a knowledge-based pairwise potential for proteins that depends on the relative position and orientation between residues. Using a minimalist representation of only three backbone atoms per residue, KORP utilizes a 6D joint probability distribution to outperform state-of-the-art statistical potentials for native structure recognition and best model selection in recent CASP and loop modeling benchmarks. Compared with the existing methods, our side-chain independent potential has a lower complexity and better efficiency. The superior accuracy and robustness of KORP represent a promising advance for protein modeling and refinement applications that require a fast but highly discriminative energy function.

Availability: http://chaconlab.org/modeling/korp
Contact: pablo@chaconlab.org
Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

The knowledge of 3D protein structure is one of the key steps for understanding biological function and ultimately pursuing protein design. However, the number of known protein structures is approximately a thousand times smaller than the number of identified protein sequences. Since these figures will probably never reconcile, the development of reliable methods for protein structure prediction and design are the current major challenges in structural bioinformatics. In this context, it is critical for model refinement, quality assessment, and conformational sampling to have an efficient and accurate energy function for the discrimination of native or near-native conformations from large decoy sets. Energy functions can be generated from either a physical or a statistical energy model. Despite the fact that physics-based potentials, including all-atom molecular mechanics force fields, are generally more accurate, knowledge-based potentials (KBPs) derived from known protein structures are a practical alternative because of their excellent balance between accuracy and speed.

Since the pioneering KBP methods that used a simple contact definition based on a distance cutoff (Betancourt and Thirumalai, 1999; Miyazawa and Jernigan, 1996; Park and Levitt, 1996; Sippl, 1990; Skolnick, et al., 2000; Tanaka and Scheraga, 1976), a considerable number of approaches have been developed (Poole and Ranganathan, 2006). Among the major improvements are the consideration of distance dependence (Gohlke and Klebe, 2001; Lu and Skolnick, 2001; Zhou and Zhou, 2002) and the inclusion of orientational terms (Bahar and Jernigan, 1996; Buchete, et al., 2004; Miyazawa and Jernigan, 2005;

Mukherjee, et al., 2005). Current KBPs, such as OPUS-PSP (Lu, et al., 2008), RW+ (Zhang and Zhang, 2010), GOAP (Zhou and Skolnick, 2011), SOAP (Dong, et al., 2013), ROTAS (Park and Saitou, 2014), ORDER_AVE (Liu, et al., 2014), ICOSA (Elhefnawy, et al., 2015), and SDFIRE (Hoque, et al., 2016), further exploit in many different flavors the distance and the orientation preferences of the pairwise contacts.

The relative position and orientation between two interacting residues or any two 3D objects can be described with six degrees of freedom (i.e., three translations plus three rotations or one distance plus five angular variables). Therefore, a six-dimensional joint probability distribution should be used to describe the relative position between residue pairs. However, it is quite challenging to recover a detailed joint distance and orientation relation from the relatively low number of known structures. To overcome the insufficient statistical data, some authors have assumed independence between the angular parameters of the potential (e.g., GOAP), and others have reduced the dimensionality of the joint probability function (e.g., 4D in ORDER_AVE).

The consideration of different coarse-grained approximations is another successful strategy in the KBP field. The ability of coarse-grained statistical potentials to perform at the level of detailed all-atom models has already been reported (Buchete, et al., 2004; Colubri, et al., 2006; Fitzgerald, et al., 2007; Melo, et al., 2002; Zhang and Kim, 2000). In principle, considering the atomic nature of the interaction should be more sensitive in distinguishing the small conformational differences. However, the errors and inaccuracies typically found in the modeled atomic coordinates can lead to severe energy distortions that may require additional costly refinement steps. Strikingly, methods with reduced representations reported competitive sensitivities and accuracies to the atomic potentials but at a reduced computational cost. The advantage of summarizing the energy terms into a few pseudoatoms increases computational efficiency while alleviating problems related to the lack of statistics and the low quality of the decoy structures.

Here we present a new coarse-grained potential for proteins defined by a 6D joint probability that only depends on the relative orientation and position of three backbone atoms. This side-chain independent potential, named KORP, is the first attempt to capture the 6D nature of the residue interactions with a reduced framework. The method has been thoroughly validated using a modern, diverse, even, and continuous distribution of decoy conformations generated by 3DRobot (Deng, et al., 2015). KORP results on these robust benchmarks and on recent CASP decoy datasets show superior performance over all other comparable potential approximations. Furthermore, since we are particularly interested in loop modeling (Chys and Chacon, 2013; Lopez-Blanco, et al., 2016), we also proved the outstanding ability of KORP to discriminate loop decoys at different loop lengths.

2 Methods

2.1 Relative position and orientation of reference frames

Our geometrical framework is equivalent to other KBPs such as GOAP (Zhou and Skolnick, 2011) but adapted to consider a single frame per interacting residues pair instead of many. As depicted in Fig. 1, we expressed the relative position and orientation of two interacting amino acids *i* and *j* with one distance and five angular parameters. The former, $\boldsymbol{r}_{ij}$, corresponds to the axis connecting the two alpha carbon (C_α) atoms. The relative orientation of each residue pair is described by five angles: θ_i, φ_i, θ_j, φ_j and ω_{ij}. The first four angles are the polar coordinates of the $\boldsymbol{r}_{ij}$ vector in the local 3D reference frame of each amino acid, defined as:

$$\begin{aligned} \boldsymbol{V}_z &= (\boldsymbol{r}_{CC\alpha} + \boldsymbol{r}_{NC\alpha}) \Big/ \left| \boldsymbol{r}_{CC\alpha} + \boldsymbol{r}_{NC\alpha} \right| \\ \boldsymbol{V}_y &= (\boldsymbol{V}_z \times \boldsymbol{r}_{NC\alpha}) \Big/ \left| \boldsymbol{V}_z \times \boldsymbol{r}_{NC\alpha} \right| \\ \boldsymbol{V}_x &= (\boldsymbol{V}_y \times \boldsymbol{V}_z) \end{aligned} \quad (2)$$

where $\boldsymbol{r}_{CC\alpha} = \boldsymbol{r}_C - \boldsymbol{r}_{C\alpha}$ and $\boldsymbol{r}_{NC\alpha} = \boldsymbol{r}_N - \boldsymbol{r}_{C\alpha}$ are vectors defined from the C_α to the carbonyl carbon (C) and nitrogen (N) atoms of the same residue, respectively. Note that none side-chain atom is needed. Finally, the torsional angle ω_{ij} describes the relative rotation of vectors $\boldsymbol{V}_{zi}$ and $\boldsymbol{V}_{zj}$ along the $\boldsymbol{r}_{ij}$ axis.

State-of-the art KBPs applied the framework to every pair of interacting atoms, on the contrary, KORP considers a single framework per interacting residues pair. Another singular difference is the full dependence of the 5 angles and the distance in our statistical potential.

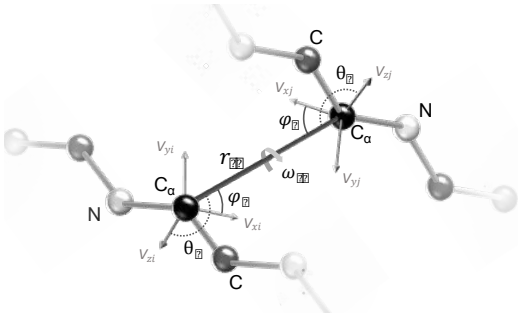

Fig. 1. Schematic view defining the relative orientation and position of two interacting residues. The relative orientation is described by two polar angles (θ and φ) for each residue *i* and *j* plus one torsional angle, ω . θ is the angle between the $\boldsymbol{r}_{ij}$ and V_z vectors, and φ is the angle between V_x and the projection of $\boldsymbol{r}_{ij}$ into the plane defined by V_x and V_y. ω_{ij} is the dihedral angle defined by the vectors $\boldsymbol{V}_{zi}$, $\boldsymbol{r}_{ij}$ and $\boldsymbol{V}_{zj}$.

2.2 Potential definition

The KORP potential was derived from known protein structures using the classical inverse Boltzmann equation:

$$E_{ij} = -RT \ln \frac{P_{ab}^{obs}\left(r_{ij}, \varphi_i, \theta_i,\ \varphi_j, \theta_j, \omega_{ij}\right) + z}{P^{ref}\left(r_{ij}, \varphi_i, \theta_i,\ \varphi_j, \theta_j, \omega_{ij}\right) + z} \quad (2)$$

where P_{ab}^{obs} is the joint probability at the given relative distance and orientation of observing two amino acids *i* and *j* of type *a* and *b,* respectively. Unlike previous approaches, KORP exploits the full 6D joint probability distribution. For example, GOAP assumes the independence between 6D variables and ORDER_AVE only considers the inter-dependence in four variables. The added *z* constant is a simple trick to prevent infinite values for very small probabilities and to improve the numerical stability for low-count statistics. A critical question that influences the performance of statistical potentials is the proper definition of a reference probability, P^{ref} . In our case, we used the classical reference state (Samudrala and Moult, 1998) that simply averages over the 20 different amino acid types to represent the reference probability regardless of residue type. Ignoring the residue chemical identity in the reference state overlooks the nonspecific interactions while enhancing specific amino acid type contributions. Finally, the energies were zero mean normalized at every distance to smooth distance-specific differences. To generate a uniform sampling of the polar angles we employed an equal-area tessellation approach (Beckers and Beckers, 2012). This method has the advantage of easily regulating

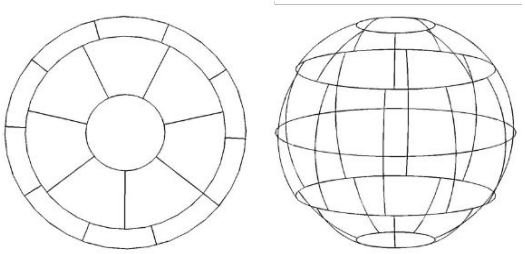

Fig. 2. Uniform angular sampling tessellation. Top and perspective views of the equal-area tessellation used in KORP. It divides the sphere surface into 36 parts,

the angular sampling coarseness while also preventing oversampling at the poles (Fig. 2).

As many other potentials, we also separated the local and nonlocal interactions. Thus, the overall energy is defined as a linear combination:

$$E = f\sum_{n<s}^{n=2} E_{ij}^{\text{local}} + \sum_{n\geq s}^{n=s} E_{ij}^{\text{non-local}} \quad (3)$$

where n is the separation in the sequence index between the i-th and j-th ($n = j - i$) residues, s is the threshold to distinguish local from nonlocal interactions and f is a weighting factor. Every local and nonlocal interaction contains the energy contribution for each of the 400 different 6D maps representing all possible interactions between the 20 canonical amino acid types. Note that P_{ab}^{obs} and P_{ba}^{obs} have subtle but significant differences because the orientation preferences between residues of type a and b do not generally commute, e.g., it is not the same to have a proline before an alanine (in sequence) as it is after.

Since our approach is computationally very efficient, we were able to scan reasonable ranges of parameters to find good results in all the modeling benchmarks described below. It is important to mention that we obtained equivalent results within a relatively wide range of parameters, so we arbitrarily selected those that gave us a potential with a reasonable granularity and file size. The distance range to consider interactions between residues was limited from 3 to 16 Å and divided into 10 bins. We divided the φ and θ space into 36 equal-area bins using the abovementioned tessellation approach. This led to an angular sampling of around 34°. The ω angle was divided into 8 bins of 45°. Nevertheless, other combinations of larger bins and lower distance limits also performed well. The interaction was considered local when residues were less than five residues apart in sequence (s = 5). We used a factor f of 1.8 for weighting the local with respect to the nonlocal contribution. Slightly better results were found for protein modeling using this distinction, but it had no effect for loop modeling. We set z to 10^{-8}.

Motivated by the results obtained with other KBPs, we also tried to increase the complexity with no success. For example, we found no improvements using more than one interaction framework per residue or considering sequence separation of the local interactions.

2.3 Decoy datasets and benchmarks

For structure modeling, the potential has been validated with modern decoy datasets generated by 3DRobot (Deng, et al., 2015) that have significantly enhanced diversity and evenness, with a continuous distribution in the root mean square deviation (RMSD) space. The 3DRobot dataset comprises 300 decoys for 200 non-redundant target proteins with a pairwise sequence identity <20%, containing 48 α, 40 β and 112 α/β single-domain proteins with lengths from 80 to 250 residues. These enhanced decoy sets have better local structures, i.e., their sidechains are well packed and the overall structures are clash free, making the native structure recognition by trivial potentials more difficult. We also evaluated the KORP performance with two classical decoy benchmarks: the Rosetta decoy set (41 proteins with 100 structure decoys each) and the I-Tasser decoy set (Zhang and Zhang, 2010) (56 proteins with 300-500 decoys each). We employ a Rosetta benchmark subset compiled by Deng (Deng, et al., 2015) in where they remove 17 bad targets whose number of decoys with RMSD<12 A˚ is fewer than 50. To enhance the evenness and diversity of these datasets, we used high-quality decoys generated with 3DRobot (Deng, et al., 2015) (herein referenced as 3DR) for Rosetta and I-Tasser target. All these datasets were downloaded from the 3DRobot website.

We also downloaded all the targets and decoys of the quality assessment category of CASP10 (Kryshtafovych, et al., 2014), CASP11 (Kryshtafovych, et al., 2015) and CASP12 (Moult, et al., 2018) from the corresponding repository at http://predictioncenter.org. Instead of partial subsets (e.g. sel20 or best150 datasets) we consider all decoy models presented in CASP contests to have the largest and most representative datasets. Moreover, we take into account only the CASP targets with at least one structural model of GDT-TS score over 0.5 to prevent any bias from targets comprised only with bad decoys. Models were trimmed to individual domains from the posted full-length targets according to the submitted domain definitions. The final datasets from CASP10, CASP11, and CASP12 contains 78, 58, and 36 target structures, respectively (see the supplemental data for the detailed list). Each target has ~320 decoys on average.

For loop modeling, we build new benchmark datasets of 6, 8, 10, and 12 residues long loops comprising 100 targets each. The loops were randomly selected from the structures included in our PISCES training set excluding any homologue structure (identity greater than 50%) used as training set in SOAP-loop approach (Dong, et al., 2013). For each loop target case, we generated 1000 geometric feasible loop decoys with our loop closure approach RCD (Chys and Chacon, 2013). Side chains were included and repacked using Rosetta. To evaluate the loop modeling performance, we employed the RMSD between the generated loop decoys and the native loop using N, C_α, C and O atoms. All the benchmarks used in this study are available for download.

2.4 Known protein structure database

The statistical energy function was extracted from 250 million contacts present in a dataset of 36851 non-redundant protein chains taken from the PISCES web server (Wang and Dunbrack Jr, 2003) and having less than 90% sequence identity, along with a resolution better than 3.0 Å and an R-factor below 0.25. We also tested a smaller subset of 50% maximum sequence identity (~50% less contacts) with slightly worse results (see Supplementary Information Fig. S5).

To avoid overtraining effects in KORP, for every benchmark described above we excluded all structures in the PISCES training set with a sequence identity ≥50% to the corresponding test structures using CD-HIT (Fu, et al., 2012). This protocol was also performed with an in-house fast version of ORDER_AVE to also avoid such effects in this method (see supplemental material Appendix B).

2.5 Model quality assessment

For model quality assessment, we adopted the following evaluation metrics commonly used in the CASP community (Kryshtafovych, et al., 2015), based on the GDT_TS score (Zemla, 2003):

(1) **N**. Number of benchmark cases where the native conformation ($\mathbf{N_N}$) or the best decoy ($\mathbf{N_D}$) have the lowest energy.

(2) **Z**. This metric measures the number of standard deviations (σ) between the energy of the native structure (E_{native}) and the mean energy of the decoys (μ): $Z_N = (\mu - E_{native})/\sigma$. In the same way, it can be defined with respect to the best decoy (closest to the native) as: $Z_D = (\mu - E_{best-decoy})/\sigma$.

(3) **Loss**. Loss of quality between the best decoy available and the lowest energy model in percentage of GDT_TS score.

(4) $\mathbf{N_{0.5}}$. Number of cases in the benchmark in where the lowest energy decoy has a GDT_TS score larger than 0.5.

(5) **Rank**. Rank of the lowest energy decoy. To facilitate comparison between benchmarks, it is expressed as the percentage of the total number of decoys for each target.

(6) **ΔAUC.** Is the difference between two areas under curves. The first curve is the number of targets with at least one decoy at the given value of GDT_TS, and the second is the number of targets in where the lowest energy decoy scores better than the given GDT_TS value (see Supplementary Information Fig. S1). To focus only on cases where at least a good model is present, this difference was restricted between 0.5 and 1.0 GDT_TS scores. As before, it is expressed as a percentage. Low percentages indicate that the lowest energy decoys are very close to the best available and vice versa.

(7) **r**. Cross-correlation coefficient (Pearson´s) between the calculated energy and the GDT_TS score.

Note that any metric based on the native structure (N_N or Z_N) is less useful in a real/blind structure prediction in where best-decoy based metrics are more adequate. Measures based on the lowest energy decoy such as $N_{0.5}$, Loss, or Rank are direct indicators of the best possible performance. Interestingly, ΔAUC integrates the ability to detect the best decoy in the range where the correct fold can be identified.

3 Results

3.1 Protein Modeling

We compared our KORP potential with six state-of-the-art KBPs for model quality assessment: RW+ (Zhang and Zhang, 2010) includes a side-chain orientation-dependent term and uses a random walk chain for the reference state; GOAP (Zhou and Skolnick, 2011) combines the atomic DFIRE potential with a six independent angular parameters potential; ICOSA (Elhefnawy, et al., 2015) combines an icosahedral tessellation of the three-dimensional interaction space with a minimalist representation of three backbone atoms per residue; ORDER_AVE (Liu, et al., 2014) uses a five atoms per residue potential with a four-dimensional joint probability distribution; VoroMQA (Olechnovic and Venclovas, 2017) uses Voronoy tessellation and interatomic distances; and finally, OPUS-DOSP (Xu, et al., 2017) is a recent update of OPUS-PSP (Lu, et al., 2008) that includes a new orientation and distance auxiliary function. The comparison was performed in two different testing scenarios: modern decoy datasets generated by 3DRobot (Deng, et al., 2015) and datasets compiled from recent prediction CASP experiments (Kryshtafovych, et al., 2015; Moult, et al., 2014; Moult, et al., 2018). Despite important advances in protein modeling, only in a fraction of the CASP targets was it possible to identify acceptable folds, i.e., models with GDT_TS scores above 0.5 (see Supplementary Information Fig. S2). To compensate the inhomogeneous nature of the CASP decoys, we first employed 3DRobot, a recent, structurally diverse benchmark with good quality structures and continuous conformational distribution between far and near-native decoys.

3.1.1 Native and near-native recognition with 3DRobot datasets

The ability to identify the native structure from decoys is used in the field as a primary test of discriminative power, but it is quite limited in a real modeling scenario, where decoy structures are quite inaccurate. For this reason, we also checked the ability to select the closest to the native model among the decoys.

As shown in Table 1, in the original 3DRobot dataset, our approach has better scores than any other. For example, KORP generally has the highest Z-scores, it is the only method that recognized 143 out of 200 of the best decoys, has the lowest Loss (1.1%), Rank (0.6%) and ΔAUC (1.1%), and obtained the second highest correlation. Notably, ORDER_AVE, that also employs a joint probability, ranks at a close distance and shows remarkably better results than GOAP, DOSP, VoroMQA, RW+, and ICOSA. In the smaller and less challenging 3DR datasets of Rosetta and I-Tasser, KORP also stands out from the other methods in the majority of metrics. Only in I-Tasser-3DR dataset, ORDER_AVE has slightly better Loss, Rank, and ΔAUC than KORP, and it has higher correlation as in the other benchmarks.

Table 1. 3DRobot datasets.

Benchmark[1]	Potential	N_N	Z_N	N_D	Z_D	Loss	$N_{0.5}$	Rank	ΔAUC	r
3DRobot (200)[2]	KORP	**193**	3.03	**143**	**2.32**	**1.1**	**200**	**0.6**	**2.1**	0.90
	RW+	5	1.23	83	1.72	6.6	198	2.6	13.0	0.86
	GOAP	131	2.02	64	1.76	4.8	199	2.0	9.8	0.90
	ICOSA	2	1.20	20	1.40	15.3	190	7.1	30.1	0.83
	OrAve	192	2.38	136	2.00	1.2	**200**	**0.6**	2.3	**0.91**
	VoroMQA	121	1.93	54	1.70	6.7	199	2.8	13.6	0.89
	DOSP	153	**3.43**	20	1.49	23.1	165	13.2	42.8	0.43
Rosetta (41) 3DR	KORP	**37**	**3.35**	31	**2.53**	**0.3**	**41**	**0.3**	**0.7**	0.88
	RW+	0	0.94	23	1.87	4.9	40	2.4	8.8	0.80
	GOAP	29	2.27	17	1.88	4.9	40	2.5	8.7	0.85
	ICOSA	0	1.37	9	1.54	14.2	38	6.0	26.8	0.80
	OrAve	**37**	2.51	**33**	2.16	1.5	**41**	0.6	2.9	**0.89**
	VoroMQA	17	1.89	9	1.82	6.2	40	3.2	11.3	0.86
	DOSP	31	3.12	3	1.47	21.6	33	11.5	40.3	0.38
I-Tasser (56) 3DR	KORP	**38**	**3.06**	**37**	**2.54**	4.5	**56**	2.1	8.9	0.85
	RW+	0	0.81	19	1.90	12.3	53	3.3	23.6	0.79
	GOAP	24	1.85	17	1.88	9.1	**56**	3.9	17.9	0.84
	ICOSA	3	1.39	3	1.54	17.6	54	6.1	34.7	0.78
	OrAve	37	2.37	**37**	2.21	**4.2**	**56**	**1.6**	**8.4**	**0.87**
	VoroMQA	20	1.90	9	1.89	10.3	**56**	3.2	20.4	0.84
	DOSP	20	2.17	3	1.17	32.1	39	19.6	57.9	0.22

[1]All benchmarks were taken from (Deng, et al., 2015) and downloaded from https://zhanglab.ccmb.med.umich.edu/3DRobot/decoys. [2]In parentheses, the number of protein test cases included in the benchmark. The evaluation metrics are explained in section 2.5. Best values are highlighted in bold.

We obtained similar results with the original Rosetta and I-Tasser benchmarks (Table 2). KORP still prevails over the other potentials and in particular with the Rosetta set. In the I-Tasser set, our approach has a better native Z-score and Loss, where the rest of the parameters are marginally better for RW+, GOAP, and ORDER_AVE. The different nature of these original data sets, their reduced number of cases, and their lack of diversity in terms of RMSD complicates the interpretation of results. However, as before, KORP showed better native and the best decoy recognition, closely followed by ORDER_AVE. It is important to

mention that KORP and ORDER_AVE are free from over-training effects since we thoughtfully removed all the homologs to the test targets of the corresponding training sets. The excellent native recognition exhibited by DOSP with the 3DRobot set or RW+ with the I-Tasser set was likely an overfitting effect compared with their performance in other datasets.

Table 2. Classical datasets.

Benchmark[1]	Potential	N_N	Z_N	N_D	Z_D	Loss	$N_{0.5}$	Rank	ΔAUC	r
Rosetta	KORP	**24**	**3.30**	**5**	0.94	**8.0**	**26**	25.1	**27.8**	0.41
(41)	RW+	15	1.62	2	0.74	12.9	23	29.5	48.5	0.37
	GOAP	24	2.51	3	**1.03**	11.0	22	20.7	39.3	0.43
	ICOSA	19	2.14	1	0.74	11.4	24	27.3	42.0	0.39
	OrAve	21	2.45	2	0.93	10.6	**26**	**24.9**	38.7	**0.43**
	VoroMQA	19	2.35	4	0.91	11.3	25	25.8	37.9	0.39
	DOSP	6	0.85	0	0.24	14.6	21	40.0	47.7	0.05
I-Tasser	KORP	52	**6.60**	1	0.61	**9.5**	42	33.3	35.0	0.47
(56)	RW+	**56**	6.08	1	0.62	9.6	**43**	27.8	**34.7**	0.50
	GOAP	45	5.79	**2**	0.70	10.5	39	29.8	38.2	0.48
	ICOSA	28	2.29	1	0.72	11.0	38	29.8	40.6	0.50
	OrAve	52	6.38	1	**0.92**	11.3	37	**23.5**	37.7	**0.55**
	VoroMQA	50	5.59	0	0.54	11.4	39	35.0	40.5	0.46
	DOSP	50	5.55	0	0.34	14.2	35	39.8	51.8	0.20

[1]See section 2.5 for metrics definition. Best values are highlighted in bold.

3.1.2 Performance on recent CASP datasets.

The CASP10 (Moult, et al., 2014), CASP11 (Kryshtafovych, et al., 2015), and CASP12 (Moult, et al., 2018) datasets offer more realistic comparative tests. To have the largest and most representative datasets, all the server predictions of the quality assessment experiments were included (see section 2.3 for details).

As seen in Table 3, KORP again shows more top results in majority of the metrics (highlighted in black), but if not ranks second. KORP obtained generally the best values for Loss and ΔAUC, pointing out that decoys selected by our approximation are on average better (closest to the best possible) than any other method. Despite the superiority of our approach, GOAP, ORDER_AVE, and VoroMQA performed well in all the CASP datasets. ORDER_AVE have always the highest correlation values and performed particularly well in CASP10 dataset. Interestingly, GOAP had the highest values in terms of best decoy Z-score.

As a blind test suggested by the reviewers, we check the relative performance with the just released CASP 13 results (December 2018). As it can be seen, KORP obtains better results than the other methods, validating the superior performance of our approach.

It is worth noting that only for ORDER_AVE and KORP tests we removed the CASP native structures and close homologs from the corresponding training sets. Thus, we cannot discard that the other methods suffer over-training problems.

3.1.3 Efficiency

Comparatively, KORP is the faster approach. For example, processing the 3DRobot 200 test cases with 300 decoys each takes ~6 minutes in a standard Linux PC, including the I/O, the most time-consuming part. Notably. This computation time can be reduced by a half using only N, C_α, and C backbone atoms as input. The majority of the tested potentials have an unexpectedly slow performance. GOAP and the original ORDER_AVE approach each take several hours to process the 3DRobot data set. KORP's lower complexity is behind its superior efficiency, it only requires 3 atoms per residue compared with existing methods that typically use more than a hundred different types of atoms (e.g. GOAP). Even ORDER_AVE, that uses a simpler 4D joint probability distribution, requires five backbone-atoms (including C_β) to compute different local and nonlocal interactions. By contrast, KORP only measures the distance between close C_α atoms and, if they are less than 16 Å apart, computes the five angles per residue-residue contact.

Table 3. CASP datasets including all the decoys.

Benchmark[1]	Potential	N_N	Z_N	N_D	Z_D	Loss	$N_{0.5}$	Rank	ΔAUC	r
CASP10	KORP	53	**2.41**	6	**1.15**	**5.3**	**71**	**14.8**	**15.6**	0.69
(78)[2]	RW+	33	1.40	5	0.85	11.5	67	21.3	33.9	0.54
	GOAP	48	1.98	6	**1.15**	7.5	72	17.4	24.6	0.62
	ICOSA	7	0.86	1	0.83	14.7	65	19.4	42.0	**0.72**
	OrAve	**58**	2.00	**10**	1.13	6.0	**71**	**14.8**	19.5	**0.72**
	VoroMQA	48	1.93	3	1.14	6.2	73	17.8	20.7	**0.72**
	DOSP	27	1.16	1	0.33	17.1	57	33.6	39.2	0.04
CASP11	KORP	**42**	**2.57**	**9**	1.31	7.6	46	**9.9**	**19.6**	0.72
(58)[2]	RW+	20	1.33	2	0.92	12.7	44	18.7	33.3	0.51
	GOAP	35	2.08	4	**1.39**	**6.9**	**50**	11.4	19.8	0.65
	ICOSA	5	0.89	1	0.78	14.5	40	21.7	38.6	0.69
	OrAve	38	2.03	7	1.24	8.9	44	10.7	24.3	**0.76**
	VoroMQA	38	2.18	7	1.29	9.1	45	13.8	27.1	0.73
	DOSP	22	1.24	1	0.45	19.5	35	29.4	42.9	0.07
CASP12	KORP	**20**	**2.07**	**2**	1.30	**8.9**	**28**	11.9	27.8	0.77
(32)[2]	RW+	8	1.17	1	1.01	12.3	26	16.5	37.7	0.68
	GOAP	12	1.58	1	**1.33**	10.0	26	12.5	**19.0**	0.71
	ICOSA	2	0.92	0	0.81	20.3	21	23.7	53.4	0.73
	OrAve	17	1.75	**2**	1.27	10.6	26	**11.2**	32.5	**0.80**
	VoroMQA	13	1.49	1	1.29	11.7	25	11.7	37.8	0.77
	DOSP	10	1.22	1	0.55	18.2	21	26.9	47.8	0.15
CASP13	KORP	**14**	**1.97**	**2**	1.11	**8.1**	**21**	**11.1**	**24.6**	0.74
(27)[2]	RW+	4	0.97	0	0.91	12.1	20	15.6	37.9	0.70
	GOAP	5	1.39	1	1.13	14.2	17	12.6	37.6	0.71
	ICOSA	1	0.71	0	0.77	19.7	18	24.3	58.8	0.72
	OrAve	10	1.53	1	1.10	11.6	19	11.4	32.4	0.77
	VoroMQA	8	1.35	1	**1.14**	13.8	18	12.8	40.5	**0.78**
	DOSP	11	1.17	0	0.24	21.7	15	35.3	58.0	0.15

[1]See Table 2. [2]Targets without any decoy >0.5 GDT-TS score were excluded.

3.2 Loop Modeling

In the absence of recent benchmarks to assess the KBPs discriminative power in a loop-modeling scenario, we built datasets of 6, 8, 10, and 12 residues long loops, each of them comprising 100 randomly selected targets. For each target, we generated a random ensemble of 1000 geometric feasible loop decoys using our loop-closure algorithm RCD (Chys and Chacon, 2013) and computed their RMSD from the native loop (see methods for details). These loop ensembles were ranked by several KBP potentials to assess their relative ability to discriminate the closest to native models. Since the exponential growth of possible conformations with loop size makes incrementally challenging to find close-to-native loops from pure geometrical sampling, we ran independent tests with each dataset. We centered our comparison with the best available KBP developed specifically for loops, i.e. SOAP-loop (Dong, et al., 2013). This potential outperformed classical potential for loop decoy discrimination, and for example, is one of the key components in current loop modeling programs such as Sphinx (Marks, et al., 2017). We extended the analysis to ORDER_AVE because it was one of the best performants in the modeling scenario, as well as, ICOSA

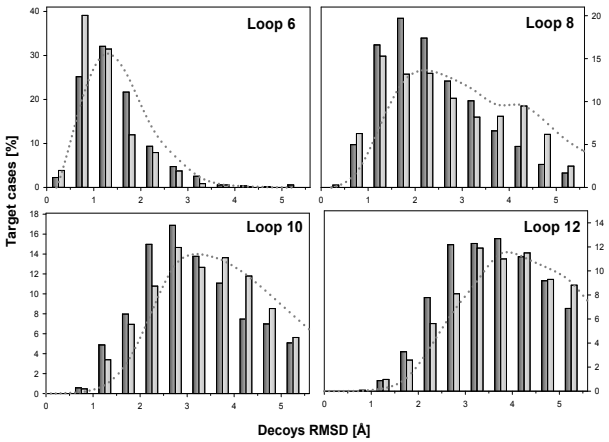

Fig. 3. Comparative RMSD distributions observed for the top 10 lower energy loop decoys. RMSD distributions found in the 1% of the KORP (dark gray) and SOAP-loop (light gray) lowest energy decoys. The dashed line corresponds to

because we use it in our loop-modeling server (Lopez-Blanco, et al., 2016) as energy filter prior final Rosetta minimization.

Table 4 shows the backbone RMSD for the closest-to-the-native model of the top 1, 10, and 100 ranked decoys. As it can be seen, KORP retrieved better decoys in all the cases except for six residue long loops in where SOAP-loop and Rosetta are superior. With larger loops (≥ 8) KORP obtains approximately the same RMSDs with 10 times fewer decoys than ICOSA, and approximately requires 2-4 folds fewer decoys than SOAP-loop or ORDER_AVE (see Fig. S3). ORDER_AVE performance is worse (0.2~0.4 Å higher RMSDs) than KORP but still competitive with SOAP-loop at larger size loops. The improved ability of KORP to discriminate near-native loop decoys is further illustrated in Figure 3 by the corresponding RMSD distributions observed in the top 10 decoy loops with lower energy (i.e., 1% of the ensemble). Comparatively, there is a greater enrichment of lower RMSD decoys with respect to the original distribution using KORP than with SOAP-loop (see also Supplementary Information Fig. S4). Only at RMSDs very close to the native (below ~1 Å) and in the shorter cases (6 and 8 residues) SOAP-loop obtains better results than KORP. The better relative recognition between decoy structures very close to native is likely related to the more detailed atomic nature of SOAP-loop potential. By the contrary, the coarse-gained nature of KORP allowed a wider discrimination range and better selection of the best models from distant decoys.

Table 4. Loops modeling.

		Loop length			
	Potential	6	8	10	12
Top 1	KORP	1.44[1]	**2.18**	**3.12**	**4.11**
	SOAP-loop	**1.05**	2.61	3.49	4.59
	OrAve	1.69	2.78	3.51	4.49
	ICOSA	2.65	3.85	4.72	5.68
Top 10	KORP	0.88	**1.35**	**1.89**	**2.58**
	SOAP-loop	**0.69**	1.71	2.11	2.98
	OrAve	1.04	1.72	2.02	2.78
	ICOSA	1.67	2.27	2.67	3.61
Top 100	KORP	0.62	**1.05**	**1.48**	**2.02**
	SOAP-loop	**0.56**	1.10	1.68	2.24
	OrAve	0.67	1.15	1.52	2.07
	ICOSA	0.86	1.38	1.67	2.60

[1] Average lowest backbone RMSD of the indicated top scoring decoys.

3.2.1 Efficiency

Another advantage of our approach is the efficiency. For example, KORP takes less than a minute to process all the 100,000 loops of the 12 residue long dataset, whereas SOAP-loop takes many hours on our Linux PC. However, this difference is much higher than expected because the available SOAP-Loop python script is far from being optimal for handling multiple loops. In any case, SOAP-loop depends on both solvent accessibility calculations and the distances and orientations between all heavy atoms; this complexity is clearly higher than our side-chain independent approach, which only uses a single interaction per residue.

4 Discussion

We present a novel potential grounded on a residue coarse-grained representation and a 6D joint probability distribution that comprehensively considers the relative orientation and position of interacting residues. KORP has shown a robust performance and consistently improved recognition of the best model for protein models and loops. Despite its simplicity, this side-chain-independent potential shows a better discrimination power than the state-of-the-art and complex atom-based potentials in practically all the tested conditions. Only in the shorter loops cases at RMSDs below ~1Å SOAP-loop obtains better results than KORP. More importantly, our approach is comparatively much better selecting best loop decoys from distant loop conformations. The adoption of 6D joint probability distribution is the main cause of the KORP improved performance. We obtained better results than GOAP that uses an equivalent 6D framework but assuming the independence of the angular parameters. Moreover, KORP improved the performance of ORDER_AVE using two extra dimensions in the joint probability. Nonetheless, the accurate estimation of the joint distributions is limited by the coverage of the protein structure database, and thus, the future use of updated datasets or redundancy-weighting strategies (Yanover, et al., 2014) could be positive.

Thanks to the excellent trade-off between accuracy and simplicity, KORP is a very interesting alternative for protein modeling and refinement applications that require an efficient and discriminative energy. For example, the superior ability of KORP for model ranking and selection would further enhance the current sophisticated quality assessment meta-methods that integrate many different scores with new machine-learning approaches (Cao, et al., 2017; Jing, et al., 2016; Uziela, et al., 2017). Future versions of our loop-modeling server RCD+ (Lopez-Blanco, et al., 2016) will exploit the wider discrimination range of KORP to drastically reduce the number of loop candidates to be further refined, in particular in the more challenging longer loop cases. For example, we showed that it is possible to obtain similar near-native loops with tenfold fewer decoys than the KBP implemented in RCD+ and two or three folds fewer decoys than other state-of-the-art approaches.

The simultaneous energy evaluation of decoy models or loops is well suited for parallelization. Thus, future plans for improvement include parallelization and optimization. Despite the limited structural data available, we are also interested in exploring the performance of our knowledge-based 6D potential for nucleic acid structure prediction (Miao, et al., 2017) and protein-protein docking scoring (Krueger, et al., 2014; Ramirez-Aportela, et al., 2016).

Acknowledgements

We thank Prof. Andriy Kryshtafovych for guiding us in the CASP data collection and for providing the trimmed-to-domains models, and Prof. Tong Wang for providing original code of ORDER_AVE.

Funding

This work has been supported by the Spanish grant BFU2016-76220-P.
Conflict of Interest: none declared.

References

Bahar, I. and Jernigan, R.L. Coordination geometry of nonbonded residues in globular proteins. *Folding and Design* 1996;1(5):357-370.

Beckers, B. and Beckers, P. A general rule for disk and hemisphere partition into equal-area cells. *Computational Geometry: Theory and Applications* 2012;45(7):275-283.

Betancourt, M.R. and Thirumalai, D. Pair potentials for protein folding: Choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Science* 1999;8(2):361-369.

Buchete, N.V., Straub, J.E. and Thirumalai, D. Orientational potentials extracted from protein structures improve native fold recognition. *Protein Sci* 2004;13(4):862-874.

Cao, R.*, et al.* QAcon: Single model quality assessment using protein structural and contact information with machine learning techniques. *Bioinformatics* 2017;33(4):586-588.

Chys, P. and Chacon, P. Random Coordinate Descent with Spinor-matrices and Geometric Filters for Efficient Loop Closure. *Journal of Chemical Theory and Computation* 2013;9(3):1821-1829.

Colubri, A.*, et al.* Minimalist Representations and the Importance of Nearest Neighbor Effects in Protein Folding Simulations. *Journal of Molecular Biology* 2006;363(4):835-857.

Deng, H., Jia, Y. and Zhang, Y. 3DRobot: Automated generation of diverse and well-packed protein structure decoys. *Bioinformatics* 2015;32(3):378-387.

Dong, G.Q.*, et al.* Optimized atomic statistical potentials: Assessment of protein interfaces and loops. *Bioinformatics* 2013;29(24):3158-3166.

Elhefnawy, W.*, et al.* ICOSA: A Distance-Dependent, Orientation-Specific Coarse-Grained Contact Potential for Protein Structure Modeling. *Journal of Molecular Biology* 2015;427(15):2562-2576.

Fitzgerald, J.E.*, et al.* Reduced Cβ statistical potentials can outperform all-atom potentials in decoy identification. *Protein Science* 2007;16(10):2123-2139.

Fu, L.*, et al.* CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;28(23):3150-3152.

Gohlke, H. and Klebe, G. Statistical potentials and scoring functions applied to protein-ligand binding. *Current Opinion in Structural Biology* 2001;11(2):231-235.

Hoque, M.T.*, et al.* SDFIRE: Sequence-specific statistical energy function for protein structure prediction by decoy selections. *Journal of Computational Chemistry* 2016;37(12):1119-1124.

Jing, X.*, et al.* Sorting protein decoys by machine-learning-to-rank. *Scientific Reports* 2016;6.

Krueger, D.M.*, et al.* DrugScore(PPI) Knowledge-Based Potentials Used as Scoring and Objective Function in Protein-Protein Docking. *Plos One* 2014;9(2).

Kryshtafovych, A.*, et al.* Assessment of the assessment: Evaluation of the model quality estimates in CASP10. *Proteins: Structure, Function and Bioinformatics* 2014;82:112-126.

Kryshtafovych, A.*, et al.* Methods of model accuracy estimation can help selecting the best models from decoy sets: Assessment of model accuracy estimations in CASP11. *Proteins: Structure, Function and Bioinformatics* 2015;84:349-359.

Liu, Y., Zeng, J. and Gong, H. Improving the orientation-dependent statistical potential using a reference state. *Proteins: Structure, Function and Bioinformatics* 2014;82(10):2383-2393.

Lopez-Blanco, J.R.*, et al.* RCD+: Fast loop modeling server. *Nucleic Acids Res* 2016;44(W1):395-400.

Lu, H. and Skolnick, J. A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins: Structure, Function and Genetics* 2001;44(3):223-232.

Lu, M., Dousis, A.D. and Ma, J. OPUS-PSP: An Orientation-dependent Statistical All-atom Potential Derived from Side-chain Packing. *Journal of Molecular Biology* 2008;376(1):288-301.

Marks, C.*, et al.* Sphinx: merging knowledge-based and ab initio approaches to improve protein loop prediction. *Bioinformatics* 2017;33(9):1346-1353.

Melo, F., Sánchez, R. and Sali, A. Statistical potentials for fold assessment. *Protein Science* 2002;11(2):430-448.

Miao, Z.*, et al.* RNA-Puzzles Round III: 3D RNA structure prediction of five riboswitches and one ribozyme. *RNA* 2017;23(5):655-672.

Miyazawa, S. and Jernigan, R.L. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *Journal of Molecular Biology* 1996;256(3):623-644.

Miyazawa, S. and Jernigan, R.L. How effective for fold recognition is a potential of mean force that includes relative orientations between contacting residues in proteins? *Journal of Chemical Physics* 2005;122(2).

Moult, J.*, et al.* Critical assessment of methods of protein structure prediction (CASP) - round x. *Proteins: Structure, Function and Bioinformatics* 2014;82:1-6.

Moult, J.*, et al.* Critical assessment of methods of protein structure prediction (CASP)-Round XII. *Proteins* 2018;86 Suppl 1:7-15.

Mukherjee, A., Bhimalapuram, P. and Bagchi, B. Orientation-dependent potential of mean force for protein folding. *Journal of Chemical Physics* 2005;123(1).

Olechnovic, K. and Venclovas, C. VoroMQA: Assessment of protein structure quality using interatomic contact areas. *Proteins* 2017;85(6):1131-1145.

Park, B. and Levitt, M. Energy functions that discriminate X-ray and near-native folds from well-constructed decoys. *Journal of Molecular Biology* 1996;258(2):367-392.

Park, J. and Saitou, K. ROTAS: A rotamer-dependent, atomic statistical potential for assessment and prediction of protein structures. *BMC Bioinformatics* 2014;15(1):16.

Poole, A.M. and Ranganathan, R. Knowledge-based potentials in protein design. *Current Opinion in Structural Biology* 2006;16(4):508-513.

Ramirez-Aportela, E., Ramon Lopez-Blanco, J. and Chacon, P. FRODOCK 2.0: fast protein-protein docking server. *Bioinformatics* 2016;32(15):2386-2388.

Samudrala, R. and Moult, J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *Journal of Molecular Biology* 1998;275(5):895-916.

Sippl, M.J. Calculation of conformational ensembles from potentials of mena force. An approach to the knowledge-based prediction of local structures in globular proteins. *Journal of Molecular Biology* 1990;213(4):859-883.

Skolnick, J., Kolinski, A. and Ortiz, A. Derivation of protein-specific pair potentials based on weak sequence fragment similarity. *Proteins: Structure, Function and Genetics* 2000;38(1):3-16.

Tanaka, S. and Scheraga, H.A. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules* 1976;9(6):945-950.

Uziela, K.*, et al.* ProQ3D: Improved model quality assessments using deep learning. *Bioinformatics* 2017;33(10):1578-1580.

Wang, G. and Dunbrack Jr, R.L. PISCES: A protein sequence culling server. *Bioinformatics* 2003;19(12):1589-1591.

Xu, G.*, et al.* OPUS-DOSP: A Distance- and Orientation-Dependent All-Atom Potential Derived from Side-Chain Packing. *J Mol Biol* 2017;429(20):3113-3120.

Yanover, C.*, et al.* Redundancy-weighting for better inference of protein structural features. *Bioinformatics* 2014;30(16):2295-2301.

Zemla, A. LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Research* 2003;31(13):3370-3374.

Zhang, C. and Kim, S.H. Environment-dependent residue contact energies for proteins. *Proceedings of the National Academy of Sciences of the United States of America* 2000;97(6):2550-2555.

Zhang, J. and Zhang, Y. A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. *PLoS One* 2010;5(10):e15386.

Zhou, H. and Skolnick, J. GOAP: A generalized orientation-dependent, all-atom statistical potential for protein structure prediction. *Biophysical Journal* 2011;101(8):2043-2052.

Zhou, H. and Zhou, Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Science* 2002;11(11):2714-2726.