HAL
open science

A geometric knowledge-based coarse-grained scoring potential for structure prediction evaluation

Sebastien Loriot, Frédéric Cazals, Michael Levitt, Julie Bernauer

To cite this version:

Sebastien Loriot, Frédéric Cazals, Michael Levitt, Julie Bernauer. A geometric knowledge-based coarse-grained scoring potential for structure prediction evaluation. Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM), Société Française de Bioinformatique, Jul 2009, Nantes, France. inria-00429607

HAL Id: inria-00429607

https://inria.hal.science/inria-00429607v1

Submitted on 3 Nov 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A geometric knowledge-based coarse-grained scoring potential for structure prediction evaluation

Sébastien Loriot[1], Frédéric Cazals[1], Michael Levitt[2], Julie Bernauer[1,2] ⋆

[1] Algorithms, Biology, Structure project-team, INRIA Sophia Antipolis
2004 route des Lucioles, BP 93, 06902 Sophia-Antipolis, France
firstname.lastname@sophia.inria.fr
[2] Department of Structural Biology, Stanford University School of Medicine
Stanford, CA 94305-5126, USA
michael.levitt@stanford.edu

Abstract: *Knowledge-based protein folding potentials have proven successful in the recent years. Based on statistics of observed interatomic distances, they generally encode pairwise contact information. In this study we present a method that derives multi-body contact potentials from measurements of surface areas using coarse-grained protein models. The measurements are made using a newly implemented geometric construction: the arrangement of circles on a sphere. This construction allows the definition of residue covering areas which are used as parameters to build functions able to distinguish native structures from decoys. These functions, encoding up to 5-body contacts are evaluated on a reference set of 66 structures and its 45000 decoys, and also on the often used* lattice_ssfit *set from the* decoys'R *us database. We show that the most relevant information for discrimination resides in 2- and 3-body contacts. The potentials we have obtained can be used for evaluation of putative structural models; they could also lead to different types of structure refinement techniques that use multi-body interactions.*

Keywords: knowledge-based potential, structure prediction and refinement, spherical arrangements, surface area, coarse-grained model.

1 Introduction

Amongst the forces driving protein folding, solvent effects and hydrophobic interactions are known to play the greatest role [2]. Calculation of the solvent accessible surface area has given important insights to estimate solvation energies [10,13]: methods that estimate solvation energies from surface area have proven useful for physics-based potentials [8,9]. Knowledge-based potentials, built from structures that are known to be stable in solution are expected to take solvation effects into account. These potentials are generally derived from distance measurements in known protein structures. For example, comparing the distribution of distances between two hydrophobic residues and that between a hydrophobic residue and a hydrophilic residue, shows that hydrophobic residues minimize their solvent contact.

The theoretical basis of such knowledge-based potentials have been questioned [3] but they often have proven to be as successful as physics-based potentials [12,15,19,22,24,25]. Demonstrating the validity of knowledge-based potentials has become easier with the availability of large, and good-quality

⋆ The authors thank the France-Stanford Center for Interdisciplinary Studies and the INRIA Équipes Associées program for funding, and the NSF award CNS-0619926 for computer resources.

protein decoy datasets [20,25], partly triggered by the CASP experiment (http://predictioncenter.org/). Attempts to derive knowledge-based potentials from more precise definitions, such as the Voronoi tessellation procedure, are as efficient as distance-based techniques [17]. The contacts obtained using this type of procedure often give a sharper signal, providing a more accurate description that leads to a better performing potential or scoring function [4].
Although fast and accurate accessible surface area calculations [1,5] and two- and four-body potentials [11,18,16,21] have been developed, none has addressed the problem of multi-body contact area. Recent improvements in computations of spherical arrangements give quick access to the detailed buried areas of a sphere intersected by other spheres. This has allowed us to derive potentials from these surface area computations, and consider different terms that range from accessible surface area to multi-body contacts. We show that these potentials that use an accurate description of local environments, provide a good discrimination between native/near-native structures and decoys.

2 Material and Methods

Geometric Construction. Given a set of $n+1$ spheres $S_{i,i=0...n}$ in 3-dimensional space, we consider a tuple of $k+1$ pairwise intersecting spheres $S_{i_0} \dots S_{i_k}$, and such that the volume defined by the intersection of the $k+1$ corresponding balls is non empty and is bounded by exactly $k+1$ spherical caps. For each sphere of that tuple, e.g. S_{i_l}, the part of S_{i_l} contained in all other spheres of the tuple defines a spherical cap of order k denoted $O_k(S_{i_l}, \{S_{i_1} \dots S_{i_{l-1}}, S_{i_{l+1}} \dots S_{i_k}\})$. A $(k+1)$-tuple then defines $k+1$ spherical caps of order k. The order 0 spherical cap of S_i, $O_0(S_i)$, is S_i exposed surface. The measures used in this study are the areas of the O_k surfaces. Figure 1 shows the 3 sphere case (n=2) with surfaces of order 0, 1 and 2 on S_0. An elaborate strategy to compute all $O_k(S_j, \{S_{j_1}, \dots, S_{j_k}\})$ consists in retrieving intersection pairs of spheres first, and then computing the surface arrangements for each sphere [6]. The complexity of the arrangement construction is $\mathcal{O}((n+p)\log n)$, n being the number of spheres and p the number of intersections among these spheres. Our robust implementation is based on the 3D Spherical Kernel of CGAL (Computational Geometric Algorithm Library) [7]. The structure of a protein can be described at atomic or coarse

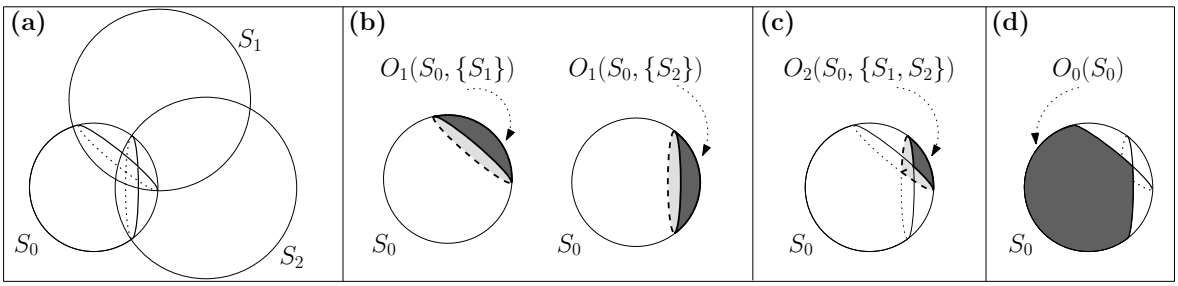

Figure 1. Definition of surface orders. (a) Spheres S_1 and S_2 intersect S_0; (b) Order 1: two surfaces of order 1 are defined by the intersections of both S_1 and S_2 on S_0; (c) Order 2: a surface area of order 2 obtained by the intersection of S_1 and S_2 in S_0; (d) Order 0: the exposed surface of S_0.

grained level using this construction. For an atomic description, O_0 is a description of the exposed surface of the protein, being either the Van der Waals surface when using the Van der Waals atomic radii or the solvent accessible surface when increasing those radii by the radius of a solvent molecule (usually 1.4Å). In this study, we use a coarse-grained representation with one sphere per residue. The center of the sphere is taken as the heavy atom closest to the center of mass of the side chain

including the Cα atom. The radii of the residues were taken from Levitt [14] and increased by 3.5Å in an arbitrary way (trial and error procedure). This value is expected to capture short and mid range interactions. For practical reasons, we limited our study to the areas of the surfaces of order up to 4 (interaction up to 5-sphere). Here the computation of the arrangement and the measure of surface areas takes an average of 10 seconds per protein structure (including all orders).

Datasets. To derive the scoring functions and assess their performance, we need a protein structure set and a set of decoys. High quality of both sets is essential if we are to obtain good quality potentials and scoring functions. For the scoring function construction, we used the dataset from Summa et al.[24] initially designed for refinement procedures. We used a subset of 66 structures corresponding to non-truncated structures, and 729 decoys for each of these structures. To assess the quality of the scoring function construction we performed a 6-fold validation in each function setting using the Summa et al. dataset. We also assessed the scoring function performance on the *lattice_ssfit* dataset from the *decoys'R us* database [25] containing eight proteins, with 2000 decoys for each.

Parameters. For each residue in a structure, the areas of all surface of order 0, 1, 2, 3 and 4 are computed. Both the value of the surface area of the spherical cap and its proportion relative to the total surface area of the sphere are used. To reduce the number of descriptors and see the influence of the residue physical and chemical properties, two types of binning were performed. The first one contains the 20 amino acids binned in 2 groups: (AGCTVILMFWY), (PSNQDERKH) and the second one the amino-acids binned in 6 groups: (FWY), (ILMV), (HKR), (DE), (NQ), (ACGPST).
For each surface of order i, with k amino acid types the number of descriptors is $N_{i,k} = k\binom{i+k-1}{i}$. Considering the large number of descriptors when the order is high, we limited our analysis to order 4, leading to no more than 756 descriptors for the intersection of 5 spheres when using 6 residue types.

Building and Evaluating the Scoring Function. For each descriptor, each value of the surface area (or its relative proportion) is measured. This is the set of observed values obs. Following the RAPDF strategy [22], the ensemble of all the measures can be defined as the reference state ref. A knowledge-based potential function can then be derived by: $E = -kT \log\left(\frac{p_{obs}}{p_{ref}}\right)$. In what follows, we will only consider the potential in its reduced form $\mathcal{S}$, with $\mathcal{S} = -\log\left(\frac{p_{obs}}{p_{ref}}\right)$.
In contrast to the usual knowledge-based potential construction, p_{obs} and p_{ref} are not obtained with a specific binning size. A kernel density estimation is performed on the data using a Gaussian kernel and the Sheather and Jones bandwidth estimation technique [23]. The data are normalized and the log odds is obtained analytically. We built 5 types of potential functions (for each order i), by summing the corresponding descriptors: $\mathcal{S}_i = -\sum_{j=1}^{N_{i,k}} \log\left(\frac{p_{obs_j}}{p_{ref}}\right)$ with $i \in [\![0,4]\!]$. One questionable approximation is whether the influence of a residue on another has the same weight independently of its relative position. This normalization and the reference state issue have been widely discussed [26]. We kept the simpler model for practical reasons. Also due to the difference of amplitude between the different terms (see section 3), they cannot be simply added to build a combined potential function. This would require a weighting scheme not addressed here.

3 Results and Discussion

Measurements and Potential Construction. Our potential is derived from 66 structures taken from the Summa et al. dataset. When considering groups of 2 residue types, there are between 3000 and 3 million values for each term of the potential. When considering groups of 6 residue types there are between 1000 and 41000 values. Some group types are more common than others, in that hydrophobic residues have a tendency to be buried and interacting less with hydrophobic residues.

Lack of sufficient data prevent us from treating individual residue types and is why we choose a coarse grained model representation (we have 2 or 6 residue types, which is much less than the 167 atom types used in the ENCAD [24] and RAPDF [22] atomic potentials).
For each residue, the equivalent of the knowledge-based potential of mean force, can be defined for surfaces of order 0 through 4. Unlike normal Lennard-Jones interactions or potentials of mean force, our potentials are not repulsive at the origin but at high value. This corresponds to the fact that the residues represented as spheres balls cannot be too close to each other.

Function Evaluation and Native Structure Identification. To quickly evaluate the relative performance of our 5 potential functions, we used the normalized rank of the native structure, i.e. the rank of the native structure divided by the total number of structures (native and decoys) considered: the lower the value, the better the performance. Table 1 summarizes results for decoys from the Summa et al. and the *lattice_ssfit* datasets. For the Summa et al. dataset, 6-fold cross validation was performed for up to order 2 surfaces when using 6 residue groups.
Results show that the exposed surface area (order 0) is not a sufficiently strong score to be able to distinguish the native structure. For order 1 surface, which is related to inter residue distance (and would be totally equivalent if the residue radii were identical), we obtain relatively good performance. Order 2 surface, which corresponds to 3-body interactions, performs slightly less well than order 1 surface but still shows discrimination power. Orders 3 and 4 surfaces, which correspond to multi-body interactions show no discrimination power. Overall, they perform no better than random selection in the Summa et al. dataset, but for some specific structure examples, they appear to give relatively good results (see figure 2).
As may be expected, the 6 residue group functions perform better than the 2 residue group functions, indicating that more than hydrophobic effects are involved in protein structure stabilization. It was not possible to decide whether the surface area or its relative proportion perform best as they seemed to do equally well.
The overall performance is much better for the *lattice_ssfit* dataset. This is to be expected as the Summa et al. decoys were built for refinement purposes and are all near-native structures, basically ranging from 0.02 to 3 Å RMSD. It is thus difficult to select the native structure, especially when we are using a single point for each residue and omitting over 90% of the atoms to make our coarse-grained models. The *lattice_ssfit* dataset contains decoys that have a different fold from the native structure, with RMSD values ranging from 4 to 16 Å.

Summa et al.		$\mathcal{S}_0$	$\mathcal{S}_1$	$\mathcal{S}_2$	$\mathcal{S}_3$	$\mathcal{S}_4$
2 groups	area	0.438 ± 0.309	0.192 ± 0.226	0.217 ± 0.219	0.450 ± 0.296	0.486 ± 0.303
	proportion	0.427 ± 0.280	0.296 ± 0.264	0.256 ± 0.251	0.464 ± 0.297	0.486 ± 0.305
6 groups	area	0.409 ± 0.298	0.137 ± 0.166	0.205 ± 0.22	0.460 ± 0.295	0.495 ± 0.306
		(0.480 ± 0.304)	(0.190 ± 0.204)	(0.239 ± 0.237)		
	proportion	0.432 ± 0.287	0.095 ± 0.151	0.285 ± 0.279	0.505 ± 0.302	0.524 ± 0.307
		(0.474 ± 0.279)	(0.202 ± 0.229)	(0.307 ± 0.288)		
lattice_ssfit		$\mathcal{S}_0$	$\mathcal{S}_1$	$\mathcal{S}_2$	$\mathcal{S}_3$	$\mathcal{S}_4$
2 groups	area	0.166 ± 0.206	0.034 ± 0.063	0.042 ± 0.073	0.16 ± 0.200	0.228 ± 0.272
	proportion	0.212 ± 0.248	0.069 ± 0.104	0.040 ± 0.063	0.199 ± 0.230	0.210 ± 0.256
6 groups	area	0.193 ± 0.197	0.023 ± 0.034	0.034 ± 0.053	0.194 ± 0.220	0.227 ± 0.277
	proportion	0.238 ± 0.327	0.002 ± 0.002	0.071 ± 0.132	0.227 ± 0.265	0.242 ± 0.285

Table 1. Normalized ranks of the native structure on the Summa et al. and on the *lattice_ssfit* datasets. Results from the 6-fold cross-validation are in parentheses.

Decoys Evaluation and Comparison With Previous Work. The normalized ranking of the native structure does not give insight on how the potential performs on near native decoys and thus whether it could be used for structure refinement. For all the protein structures evaluated, plots of score vs. RMSD were drawn. Some examples are presented in figure 2.

The exposed (order 0) surface area sometimes indicates the best structure but cannot be used as a selection criterion. In most cases the potentials for order 1 and 2 surface are able to identify and correctly rank near native structures. This is very clear for the Summa et al. dataset: plots show funnel-like shapes (high correlation between score and RMSD), characterizing good structure ranking in a refinement setting. Interestingly for some examples, order 3 and 4 potentials also show good results on the same dataset. Due to the high RMSD values of the *lattice_ssfit* dataset, the structure selection is less clear but still representative.

To compare to a recent study from Bhattacharyay et al. [5], we computed the $Zscore$, the $logPB1$ and the $logPB10$ for the *lattice_ssfit* dataset at order 1 (see [5] for definitions). We obtained average value -1.45, -0.28 and -1.42 respectively. Overall the Bhattacharyay et al. potential performs better than our best potential (-4.06, -0.41, -1.55 respectively). This may be due to the fact that their training dataset is much larger, that they look at smaller spheres or that they normalize surface areas at a single sphere level. These parameters still have to be investigated. We also expect to get better results by combining different terms of surface orders. Our study could then be extended to the whole *decoys'R us* dataset and other decoy datasets.

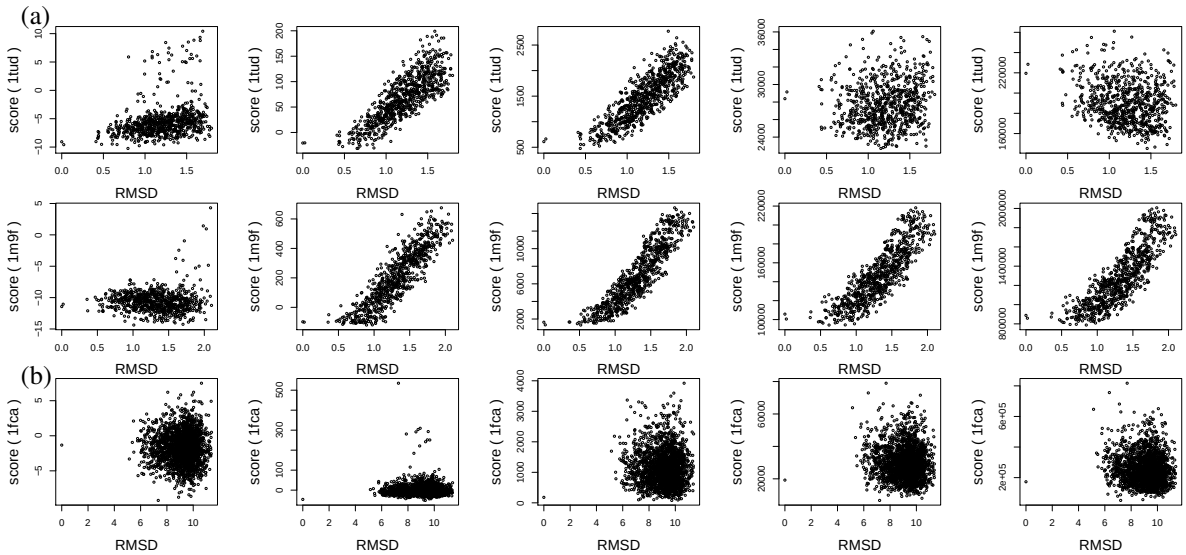

Figure 2. Examples of score vs. RMSD plots. (a) From Summa et al. dataset. (b) From the *lattice_ssfit* dataset.

4 Conclusion and Perspectives

The spherical arrangements leading to surface area measurements have proven to be a good encoding of multi-body contacts. We have shown that it is possible to derive a knowledge based potential that is able to rank correctly native and near native structures in a coarse-grained setting. Improvements over conventional very-well optimized distance-based potentials are small. Further optimization is needed, especially to combine different order scoring functions. This potential is a brand-new type, does not use distance as the primary parameter, is differentiable and could be used for minimization purposes. Combined with an atomic potential for high-resolution refinement, it could also improve structure prediction and model selection.

References

[1] Nataraj Akkiraju and Herbert Edelsbrunner. Triangulating the surface of a molecule. *Discrete Appl. Math.*, 71(1-3):5–22, 1996.

[2] R. L. Baldwin. Making a network of hydrophobic clusters. *Science*, 295(5560):1657–8, 2002.

[3] A. Ben-Naim. Statistical potentials extracted from protein structures: Are these meaningful potentials? *The Journal of Chemical Physics*, 107(9):3698–3706, 1997.

[4] J. Bernauer, J. Aze, J. Janin, and A. Poupon. A new protein-protein docking scoring function based on interface residue properties. *Bioinformatics*, 23(5):555–62, 2007.

[5] A. Bhattacharyay, A. Trovato, and F. Seno. Simple solvation potential for coarse-grained models of proteins. *Proteins*, 67(2):285–92, 2007.

[6] F. Cazals and S. Loriot. Computing the arrangement of circles on a sphere, with applications in structural biology. *Computational Geometry : Theory and Applications*, (in press), 2008.

[7] Pedro M. M. de Castro, Frédéric Cazals, Sébastien Loriot, and Monique Teillaud. Design of the CGAL 3D spherical kernel and application to arrangements of circles on a sphere. *Computational Geometry: Theory and Applications*, (in press), 2008.

[8] M. Delarue and P. Koehl. Atomic environment energies in proteins defined from statistics of accessible and contact surface areas. *J Mol Biol*, 249(3):675–90, 1995.

[9] B. N. Dominy and C. L. Brooks. Identifying native-like protein structures using physics-based potentials. *J Comput Chem*, 23(1):147–60, 2002.

[10] D. Eisenberg and A. D. McLachlan. Solvation energy in protein folding and binding. *Nature*, 319(6050):199–203, 1986.

[11] H. H. Gan, A. Tropsha, and T. Schlick. Lattice protein folding with two and four-body statistical potentials. *Proteins*, 43(2):161–74, 2001.

[12] T. Lazaridis and M. Karplus. Effective energy functions for protein structure prediction. *Curr Opin Struct Biol*, 10(2):139–45, 2000.

[13] B. Lee and F. M. Richards. The interpretation of protein structures: estimation of static accessibility. *J Mol Biol*, 55(3):379–400, 1971.

[14] M. Levitt. A simplified representation of protein conformations for rapid simulation of protein folding. *J Mol Biol*, 104(1):59–107, 1976.

[15] H. Lu and J. Skolnick. A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins*, 44(3):223–32, 2001.

[16] J. Maupetit, P. Tuffery, and P. Derreumaux. A coarse-grained protein force field for folding and structure prediction. *Proteins*, 69(2):394–408, 2007.

[17] B. J. McConkey, V. Sobolev, and M. Edelman. Discrimination of native protein structures using atom-atom contact scoring. *Proc Natl Acad Sci U S A*, 100(6):3215–20, 2003.

[18] L. A. Mirny and E. I. Shakhnovich. How to derive a protein folding potential? A new approach to an old problem. *J Mol Biol*, 264(5):1164–79, 1996.

[19] S. Miyazawa and R. L. Jernigan. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol*, 256(3):623–44, 1996.

[20] B. Park and M. Levitt. Energy functions that discriminate X-ray and near native folds from well-constructed decoys. *J Mol Biol*, 258(2):367–92, 1996.

[21] J. Qiu and R. Elber. Atomically detailed potentials to recognize native and approximate protein structures. *Proteins*, 61(1):44–55, 2005.

[22] R. Samudrala and J. Moult. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J Mol Biol*, 275(5):895–916, 1998.

[23] S. J. Sheather and M. C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(3):683–690, 1991.

[24] C. M. Summa and M. Levitt. Near-native structure refinement using in vacuo energy minimization. *Proc Natl Acad Sci U S A*, 104(9):3177–82, 2007.

[25] Y. Xia, E. S. Huang, M. Levitt, and R. Samudrala. Ab initio construction of protein tertiary structures using a hierarchical approach. *J Mol Biol*, 300(1):171–85, 2000.

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