A NEW APPROACH TO HOMOLOGY MODELING

YI HE[1], H. A. SCHERAGA[1] AND S. RACKOVSKY[1,2]

[1] *Baker Laboratory of Chemistry and Chemical Biology*
Cornell University
Ithaca, NY 14853, USA

[2] *Department of Pharmacology and Systems Therapeutics*
Icahn School of Medicine at Mount Sinai
One Gustave L Levy Place
New York, NY 10029, USA

(Paper presented at the CBSB14 Conference, May 25–27, 2014, Gdansk, Poland)

Abstract: The need to interpret experimental results led to, first, an all-atom force field, followed by a coarse-grained one. As an aid to these force fields, a new approach is introduced here to predict protein structure based on the physical properties of the amino acids. This approach includes three key components: Kidera factors describing the physical properties, Fourier transformation and UNRES coarse-grained force field simulations. Different from traditional homology modeling methods which are based on evolution, this approach is physics-based, and does not have the same weaknesses as the traditional homology modeling methods. Our results show that this approach can produce above average prediction results, and can be used as a useful tool for protein structure prediction.

Keywords: protein structure prediction, physical properties, Kidera factors, Fourier transformation, coarse-grained force field

1. Introduction

The field of protein structure prediction, currently an area of intense theoretical and computational interest, originated from an effort to understand experimental data. Protein titration and related biophysical experiments indicated that 3 of the 6 tyrosyl residues and 3 of the 11 carboxyl groups of bovine pancreatic ribonuclease A (RNase A) have abnormal $\mathrm{pK_a}$'s, suggesting that they might be involved in hydrogen bonds. With 19,800 ways to pair these possible donors and acceptors, subsequent biochemical and biophysical studies [1] provided the following unique pairing: Tyr 25…Asp 14, Tyr 92…Asp 38, and Tyr 97…Asp 83, which was then actually observed in the subsequently – determined x-ray crystal structure of RNase A. This information, together with knowledge of the location of the four disulfide bonds of the protein, provided distance constraints that motivated the development of a computational approach to predict the structure of RNase A from its amino acid sequence [2]. Further development of this computational framework [3] resulted in the ECEPP (Empirical Conformational Energy Program for Peptides) all-atom force field [4]. ECEPP was used initially to compute the structures of fibrous proteins, such as collagen and collagen models [5], and globular proteins such as the 46-residue protein A [6].

Recognizing that such a detailed force field could not be extended to globular proteins containing many more than 50 residues, a physics-based coarse-grained (UNRES, UNited RESidue) force field was developed [7]. One of its first successful applications, in CASP3, elucidated 80% (or 61 residues) of the structure of HDEA within an RMSD of 4.2 Å [8]. Following this success with UNRES, molecular dynamics was introduced, in order to compute not only protein structure, but also the folding pathways of single-chain [9] and multiple-chain [10] proteins. A recent successful application of UNRES was to a 205-residue CASP10 target with two-fold symmetry [11], providing a result that was much better than other predictions of the structure of this protein, obtained by using knowledge-based methods.

Regrettably, even coarse-grained *ab initio* computational methods are not yet sufficiently advanced to predict the structure of a protein reliably from its sequence alone. The most reliable method for predicting the structure of a target protein from its sequence remains homology modeling [12, 13]. In this approach, sequence comparison methods are used to find proteins of known structure whose sequences are similar to that of the target. The structures of those proteins are then used as starting points for modeling that of the target. For example, from a comparison of helix probability profiles from the helix-forming tendency of the naturally-occurring amino acids, it was predicted that α-lactalbumin is homologues to lysozyme [12]. This was later demonstrated [13] with ECEPP calculations. Even homology modeling, however, is not completely reliable, a fact which is reflected in two well-known phenomena:

1. Any large group of proteins known to fold to similar structures is likely to contain pairs of molecules whose sequences are not related by any known criteria [14].
2. Sequences are known to exist in which single-site mutations cause a complete change in the fold of the protein [15]. These sequences are not identifiable by current methods.

The principal current method for determining the similarity quantitatively between two sequences involves pairwise alignment of those sequences, followed by evaluation of a penalty function which accounts for the degree of correspondence between matched amino acid residues, and for the presence of insertions and deletions in one sequence relative to the other. This approach, whose use in the field is so prevalent, suffers from a number of intrinsic limitations [16]. As a result, work has been in progress for some time [17–22] on an alternative approach to the sequence comparison problem. In the present work, we combine this sequence comparison approach with the united-residue method in order to increase the reliability of homology modeling.

2. Method

This approach is based on Fourier analysis, and rests on two important ideas which are not incorporated into alignment-based approaches:

- Representation of the amino acids by orthonormal, statistically complete (but non-redundant) numerical factors based on their physical properties, rather than residue names; and
- Representation of the sequence of the protein by parameters which contain information about the *entire* sequence, rather than local information alone.

A numerical, physically-based representation of amino acids was developed by Kidera *et al.* [23, 24] who showed that all the known physical properties of the 20 amino acids can be represented by 10 property factors (shown schematically in Figure 1). These factors together carry 86% of the variance of the entire dataset of amino acid properties, and therefore the physical characteristics of each amino

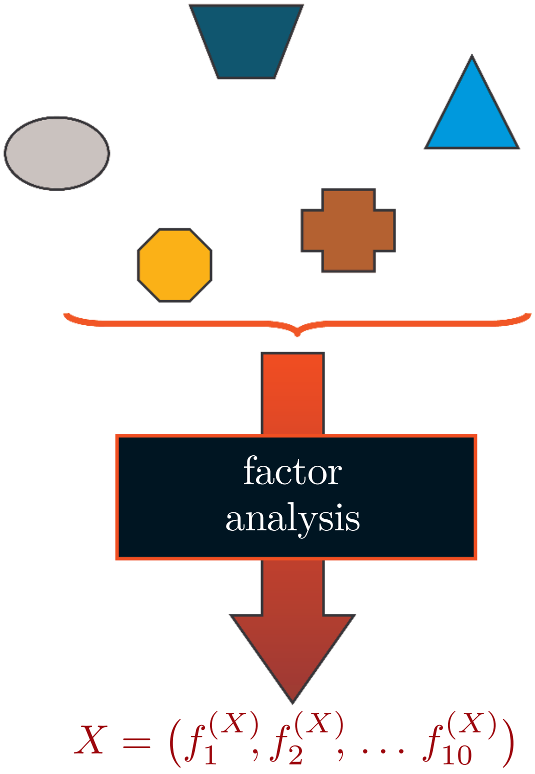
factor analysis
$X = \left(f_1^{(X)}, f_2^{(X)}, \ldots f_{10}^{(X)}\right)$

Figure 1. A schematic representation by which analysis of 188 property sets attributed to the 20 amino acids leads to a complete, orthonormal numerical representation in terms of 10 property factors [23]

acid are well represented by a 10-vector. An N-residue sequence of amino acids can therefore be represented by a set of $10N$-member numerical strings, each of which describes the values of one property factor, as a function of position, over the length of the protein.

The second idea is based on the observation by Lattman and Rose [25] that the determinants of folding must be distributed throughout the sequence of a protein. In order to properly encode these determinants, one must be able to write the sequence of the protein in terms of parameters which contain information about the entire sequence. We realize this goal by Fourier transforming the 10 numerical strings which together represent the protein sequence. The resulting (sine and cosine) Fourier coefficients $\{a_k^{(l)}\}$ are characterized by two indices – the wave number, k, and the index l $(1 \leq l \leq 10)$, which identifies the property factor whose string has been transformed as shown in Figure 2. By the definition of the discrete Fourier transform, each Fourier coefficient contains information about the entire sequence. In this connection, we note that the $k = 0$ (cosine) Fourier coefficient for the l^{th} string provides the average value of the l^{th} property factor over the sequence, but contains no information about the linear arrangement of amino acids along the chain. That information is encoded in Fourier coefficients with $k > 0$.

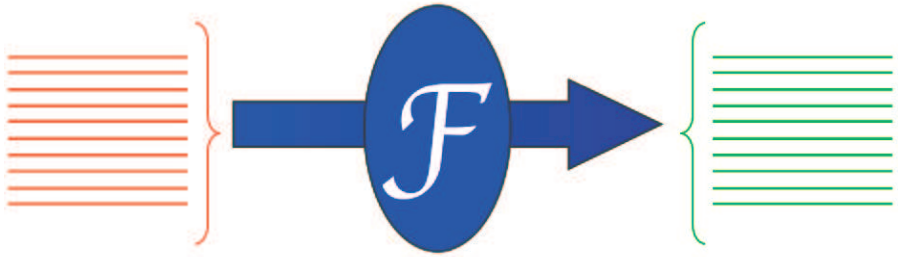

Figure 2. A schematic representation of the process by which the $10N$-member numerical strings which describe an N-residue sequence are Fourier transformed [17, 18], to give 10 strings of Fourier coefficients for each value of the wave number k

In recent work [16], we defined a distance function between sequences, based on characteristics of the Fourier sequence representation previously observed [22]. We examined the ability of that distance function, using only sequence information, to match distances between proteins based only on their structural characteristics. The inter-structure distance function used in that work was adapted from earlier work on the classification of protein structures [26, 27], and did not rely on sequence information in any way. It was shown that there is very high correlation ($R \approx 0.8$; see Figure 3) between sequence-based and structure-based distances. (This correlation, which is a *sine qua non* for correct homology detection, cannot be calculated at all within alignment-based approaches, because neither the sequence nor structure distances can be satisfactorily defined for very dissimilar molecules.) It was further shown [16] that the ability of the Fourier

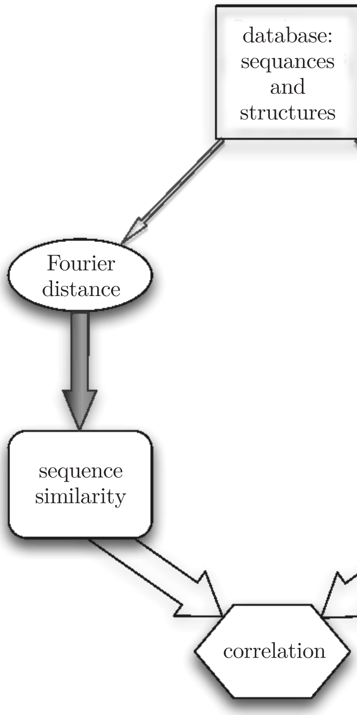

Figure 3. The process by which the correlation between inter-sequence and inter-structure distances is demonstrated [16]

distance function to correctly classify sequences is much greater than would be expected on a purely random basis, and that the algorithm performs very well in identifying homologs of actual targets of molecules of interest.

In the present work, we continue to develop the application of the Fourier representation to the identification of candidate homologs of a specified target, in conjunction with the coarse-grained UNRES force field. The Fourier representation of a target sequence was used to identify the 30 candidates closest in sequence space from the CATH database [28]. PSIPRED [29] was then used to select 5 out of those 30 candidates based on secondary structure agreement between the PSIPRED prediction of the target and the secondary structure of the candidate. Initial structures were built based on each of the 5 final candidates, and these were subsequently simulated using UNRES with MREMD. Then, a cluster analysis was carried out on the MREMD results, and the final 5 clusters selected as structural candidates for the target protein.

3. Results

The methodology was tested on a CASP8 target, T0476 (Figure 4), which is a single domain protein consisting of 108 residues. It contains three α-helices and a β-hairpin. Traditional homology modeling methods cannot predict the

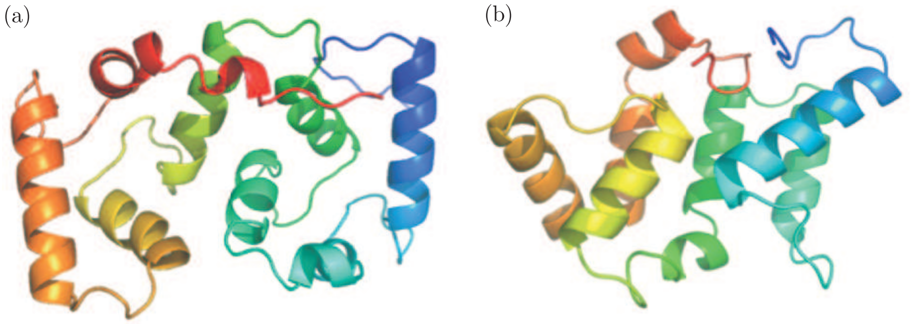

Figure 4. (a) One example of a candidate structure obtained by Fourier analysis and PSIPRED selection based on the CATH database; (b) Experimental structure of CASP8 target T0476

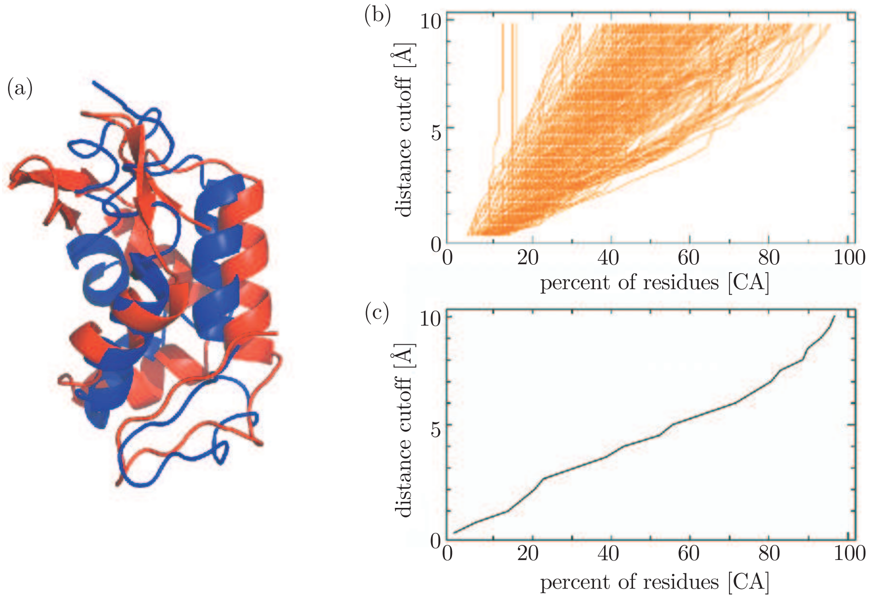

Figure 5. (a) Overlap view of the lowest rmsd structure (blue) from UNRES MREMD simulation with respect to the experimental structure (red); (b) GDT-TS plot of all models submitted by all groups for Target T0476 (The GDT-TS plots have been reproduced with permission from the CASP9 web site `www.predictioncenter.org/casp8/results.cgi`); (c) GDT-TS plot of the UNRES predicted structure shown in (a); the GDT-TS was calculated using GDT-TS server: `proteinmodel.org/AS2TS/LGA/lga.html`

correct structure of this protein, as shown in the GDT-TS plot of Figure 5 (b), released by the CASP8 website. Based on the Fourier and PSIPRED results, we selected five candidates. One candidate (among five) is shown in Figure 4 (a), and the experimental structure of T0476 is on the right (Figure 4 (b)). Starting from the structure shown in Figure 4 (a), 2,000,000 steps of MREMD simulation were performed over 32 temperatures ranging from 250 K to 500 K, and two parallel trajectories under each temperature were used in the UNRES simulation.

The lowest rmsd structure obtained in the UNRES MREMD simulation that overlapped with the experimental structure, is shown in Figure 5 (a). The GDT-TS plots of all models submitted from all groups is shown in Figure 5 (b). It can be seen that the best model from all groups can predict ∼95% of the whole chain at a 10 Å cut-off. The GDT-TS plot of the structure from our UNRES simulation, shown in Figure 5 (a), can reach ∼97% at a 10 Å cut-off.

The combination of a Fourier analysis, based on Kidera factors, and UNRES simulations can greatly reduce the computational cost of homology modeling and, at the same time, provide high quality predictions. It should be pointed out that all of this methodology is physics-based, and avoids the assumptions used in traditional homology modeling methods.

Acknowledgements

We thank the colleagues cited here for their contributions to the field. This work was supported by NIH and NSF (grant numbers GM-14312 and MCB10-19767, respectively).

References

[1] Scheraga H A 1967 *Fed. Proc.* **26** 1380
[2] Némethy G and Scheraga H A 1965 *Biopolymers* **3** 155
[3] Scheraga H A 1968 *Adv. Phys. Org. Chem.* **6** 103
[4] Momany F A, McGuire R F, Burges A W and Scheraga H A 1975 *J. Phys. Chem.* **79** 2361
[5] Miller M H and Scheraga H A 1976 *J. Polymer Sci.: Polymer Symposia* **54** 171
[6] Vila J A, Ripoll D R and Scheraga H A 2003 *Proc. Natl. Acad. Sci.* **100** 14812
[7] Liwo A, Czaplewski C, Pillardy J and Scheraga H A 2001 *J. Chem. Phys.* **115** 2323
[8] Lee J, Liwo A, Ripoll D R, Pillardy J, Saunders J A, Gibson K D and Scheraga H A 2000 *Intl. J. Quantum Chem.* **71** 90
[9] Liwo A, Khalili M and Scheraga H A 2005 *Proc. Natl. Acad. Sci.* **102** 2362
[10] Rojas A V, Liwo A and Scheraga H A 2007 *J. Phys. Chem. B* **111** 293
[11] He Y, Mozelewska M A, Krupa P, Sieradzan A K, Wirecki T K, Liwo A, Kachlishvili K, Rackovsky S, Jagiela D, Slusarz R, Czaplewski C R, Oldziej S and Scheraga H A 2013 *Proc. Natl. Acad. Sci.* **110** 14936
[12] Lewis P N and Scheraga H A 1971 *Arch. Biochem. Biophys.* **144** 584
[13] Warme P K, Momany F A, Rumball S V, Tuttle R W and Scheraga H A 1974 *Biochemistry* **13** 768
[14] Yang Y, Faraggi E, Zhao H and Zhou Y 2011 *Bioinformatics* **27** 2076
[15] Alexander P A, He Y, Chen Y, Orban J and Bryan P N 2009 *Proc. Natl. Acad. Sci.* **106** 21149
[16] Scheraga H A and Rackovsky S 2014 *Proc. Natl. Acad. Sci.* **111** 5225
[17] Rackovsky S 1998 *Proc. Natl. Acad. Sci.* **95** 8580
[18] Rackovsky S 2006 *J. Phys. Chem. B* **110** 18771
[19] Rackovsky S 2009 *Proc. Natl. Acad. Sci.* **106** 14345
[20] Rackovsky S 2010 *Proc. Natl. Acad. Sci.* **107** 8623
[21] Rackovsky S 2011 *Phys. Rev. Lett.* **106** 248101
[22] Rackovsky S 2013 *Proteins: Struct. Funct. Bioinf.* **81** 1681
[23] Kidera A, Konishi Y, Oka M, Ooi T and Scheraga H A 1985 *J. Prot. Chem.* **4** 23
[24] Kidera A, Konishi Y, Ooi T and Scheraga H A 1985 *J. Prot. Chem.* **4** 265

[25] Lattman E and Rose G 1993 *Proc. Natl. Acad. Sci.* **90** 439

[26] Rackovsky S 1990 *Polymer Prepr.* **31** 205

[27] Rackovsky S 1990 *Proteins: Struct. Func. and Genetics* **7** 378

[28] Orengo C A, Michie A D, Jones S, Jones D T, Swindells M B, Thornton J M 1997 *Structure* **5** 1093

[29] Buchan D W A, Minneci F, Nugent T C O, Bryson K, Jones D T 2013 *Nucl. Acids Res.* **41**, W340