Contact-assisted protein structure modeling by global optimization in CASP11

Keehyoung Joo,1,2 InSuk Joung,1,3 Qianyi Cheng,1,3 Sung Jong Lee,1,4 and Jooyoung Lee1,2,3*

1 Center for in Silico Protein Science, Korea Institute for Advanced Study, Seoul 130-722, Korea
2 Center for Advanced Computation, Korea Institute for Advanced Study, Seoul 130-722, Korea
3 School of Computational Sciences, Korea Institute for Advanced Study, Seoul 130-722, Korea
4 Department of Physics, University of Suwon, Hwaseong-Si, Gyeonggi-do 445-743, Korea

ABSTRACT

We have applied the conformational space annealing method to the contact-assisted protein structure modeling in CASP11. For Tp targets, where predicted residue–residue contact information was provided, the contact energy term in the form of the Lorentzian function was implemented together with the physical energy terms used in our template-free modeling of proteins. Although we observed some structural improvement of Tp models over the models predicted without the Tp information, the improvement was not substantial on average. This partly due to the inaccuracy of the provided contact information, where only about 18% of it was correct. For Ts targets, where the information of ambiguous NOE (Nuclear Overhauser Effect) restraints was provided, we formulated the modeling in terms of the two-tier optimization problem, which covers: (1) the assignment of NOE peaks and (2) the three-dimensional (3D) model generation based on the assigned NOEs. Although solving the problem in a direct manner appears to be intractable at first glance, we demonstrate through CASP11 that remarkably accurate protein 3D modeling is possible by brute force optimization of a relevant energy function. For 19 Ts targets of the average size of 224 residues, generated protein models were of about 3.6 Å Cα atom accuracy. Even greater structural improvement was observed when additional Tc contact information was provided. For 20 out of the total 24 Tc targets, we were able to generate protein structures which were better than the best model from the rest of the CASP11 groups in terms of GDT-TS.

Proteins 2016; 00:000–000.
© 2015 Wiley Periodicals, Inc.

Key words: contact-assisted modeling; global optimization; two-tier optimization; casp; sparse and ambiguous NOE restraint; protein structure modeling; NMR structure determination.

INTRODUCTION

Template-based modeling of protein three-dimensional (3D) structures has been successfully applied to various fields of protein science, but its modeling accuracy has been limited by factors such as erroneous restraint information from misaligned templates, absence of template regions, and inaccuracies of energy functions available. Recently, it has been demonstrated that more accurate protein models can be generated by incorporating information from experimental data such as sparse NMR (nuclear magnetic resonance) data,1,2 NMR chemical shift data,3,4 small angle X-ray scattering data,5,6 cryo-electron microscopy data,7,8 and residue–residue contact data.9,10

In the recent 11th CASP (Critical Assessment of Protein Structure Prediction) experiment (CASP11 in short), a new challenge called contact-assisted modeling was tested. For most of the targets in this category no obvious templates were identified. For Tp targets, predicted residue–residue contact information was collected and processed by the CASP11 organizers from the CASP11 contact prediction submissions for T0 targets. Ts targets were supplied with a set of simulated sparse and ambiguous NMR hydrogen-hydrogen contacts, and Tc targets were supplied with a set of correct residue–residue contacts.

Grant sponsor: The National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST); Grant number: 2008-0061987; Grant sponsor: The National Institute of Supercomputing and Networking/Korea Institute of Science and Technology Information with supercomputing resources including technical support; Grant number: KSC-2014-C3-01.
*Correspondence to: J. Lee Center for in Silico Protein Science, Korea Institute for Advanced Study, Seoul 130-722, Korea. E-mail: jlee@kias.re.kr
Received 15 May 2015; Revised 24 November 2015; Accepted 12 December 2015
Published online 17 December 2015 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/prot.24975
For NMR structure determination of a large protein with >160 amino-acid residues, line broadening in the NMR signal becomes an issue, and one way to overcome this is to perdeuterate the protein sample except for backbone-amide hydrogen and methyl hydrogen. From this kind of protein sample, leading protein NMR groups of today can generate NOE (nuclear Overhauser effect) signals that are sparse and ambiguous. For the Ts targets in CASP11, native crystal structures were exclusively disclosed to Gaetano Montelione’s group, and they simulated sparse and ambiguous hydrogen-hydrogen contact information, which was released to CASP11 participants as Ts targets. One motivation for experimenting with these Ts targets in CASP11 was to challenge leading computational groups working on protein structure modeling so that we may develop new methods that can handle the very high ambiguities of real NOE data. Currently, this is the major bottleneck for solving 3D structures of large proteins.

As discussed, NMR data collected from deuterated samples usually results in large spectral overlaps making proper peak assignment difficult. Here, for NMR protein structure determination, one has to deal with a combinatorial optimization problem in which the following two challenges are involved. Firstly, the ambiguous NOE peaks should be assigned to their corresponding hydrogen atom pairs and secondly, 3D structures should be constructed by satisfying as much as possible the distance restraints arranged by the given set of peak assignments. With sparse and ambiguous distance restraints, standard NMR programs for structure calculation fail to generate accurate protein 3D models and thus the protein structure determination in such a situation remains as a challenge.

Even with a set of correctly and unambiguously assigned NOE peaks (hence correct atom–atom distance restraints), generating good 3D protein models is not a trivial problem. It is a combinatorial optimization problem of finding 3D arrangements of the protein’s atoms that satisfy as much as possible assigned atom–atom distance restraints and the basic stereo-chemistry of the protein with minimal steric hindrance. When this combinatorial optimization problem is coupled with the complexity of the ambiguous peak assignment, solving it in a direct fashion (that is, modeling the protein structure by solving a two-tier optimization problem) appears to be an impossible task. In a typical Ts target, on average, a peak is assigned with more than ten possible atom–atom pairs, and the number of peaks is >1000, which gives an astronomical number of peak assignment possibilities (see Table I below).

One of the most advanced approaches to cope with the ambiguity of the NMR data is to combine it with additional information. Recently, it was demonstrated that accurate protein structure determination is possible by combining sparse NMR data with evolutionary residue–residue couplings obtained from multiple alignments of structurally related protein sequences. In contrast to this state-of-the-art approach, here, we intend to develop a computational method utilizing only the provided ambiguous contact information for modeling Ts targets. Our approach distinguishes itself in that, to the best of our knowledge, it is the first kind of a direct method solving the straightforward but daunting task of the two-tier optimization problem where the double complexity of the peak assignment and the 3D arrangement of the protein’s atoms is treated simultaneously. We will investigate the limitations and the potential of our approach.

The seemingly unattainable approach mentioned above is motivated by the successful application of conformational space annealing (CSA) to various challenging optimization problems such as the ab initio protein structure prediction, the template-based protein structure prediction, the global optimization of Lennard–Jones molecular clusters, the conformational search of protein–protein and protein–ligand interactions, community detection of a network by modularity optimization, and improved function prediction by network community detection. In this work, we will apply CSA for protein structure modeling of the Tp.

### Table I

**Analysis of Provided-Predicted-Residue-Contact Information is shown for 24 Ts Targets**

<table>
<thead>
<tr>
<th>Targets</th>
<th>N&lt;sub&gt;given&lt;/sub&gt;</th>
<th>N&lt;sub&gt;native&lt;/sub&gt;</th>
<th>N&lt;sub&gt;correct&lt;/sub&gt;</th>
<th>Correct ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tp761</td>
<td>481</td>
<td>411</td>
<td>17</td>
<td>4.1</td>
</tr>
<tr>
<td>Tp763</td>
<td>246</td>
<td>107</td>
<td>16</td>
<td>15.0</td>
</tr>
<tr>
<td>Tp767</td>
<td>412</td>
<td>347</td>
<td>82</td>
<td>23.6</td>
</tr>
<tr>
<td>Tp771</td>
<td>334</td>
<td>304</td>
<td>42</td>
<td>13.8</td>
</tr>
<tr>
<td>Tp777</td>
<td>596</td>
<td>549</td>
<td>57</td>
<td>10.4</td>
</tr>
<tr>
<td>Tp785</td>
<td>124</td>
<td>110</td>
<td>16</td>
<td>14.6</td>
</tr>
<tr>
<td>Tp794</td>
<td>543</td>
<td>543</td>
<td>244</td>
<td>44.9</td>
</tr>
<tr>
<td>Tp800</td>
<td>343</td>
<td>316</td>
<td>135</td>
<td>42.7</td>
</tr>
<tr>
<td>Tp802</td>
<td>191</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Tp804</td>
<td>294</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Tp806</td>
<td>353</td>
<td>353</td>
<td>107</td>
<td>30.3</td>
</tr>
<tr>
<td>Tp810</td>
<td>220</td>
<td>220</td>
<td>18</td>
<td>8.2</td>
</tr>
<tr>
<td>Tp814</td>
<td>685</td>
<td>673</td>
<td>142</td>
<td>21.1</td>
</tr>
<tr>
<td>Tp818</td>
<td>258</td>
<td>231</td>
<td>28</td>
<td>12.1</td>
</tr>
<tr>
<td>Tp824</td>
<td>152</td>
<td>142</td>
<td>35</td>
<td>24.7</td>
</tr>
<tr>
<td>Tp836</td>
<td>246</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Tp827</td>
<td>234</td>
<td>228</td>
<td>33</td>
<td>14.5</td>
</tr>
<tr>
<td>Tp831</td>
<td>377</td>
<td>292</td>
<td>21</td>
<td>7.2</td>
</tr>
<tr>
<td>Tp832</td>
<td>396</td>
<td>335</td>
<td>20</td>
<td>6.0</td>
</tr>
<tr>
<td>Tp835</td>
<td>668</td>
<td>642</td>
<td>163</td>
<td>25.4</td>
</tr>
<tr>
<td>Tp834</td>
<td>334</td>
<td>324</td>
<td>8</td>
<td>2.5</td>
</tr>
<tr>
<td>Tp836</td>
<td>352</td>
<td>352</td>
<td>53</td>
<td>15.1</td>
</tr>
<tr>
<td>Tp848</td>
<td>254</td>
<td>254</td>
<td>54</td>
<td>21.3</td>
</tr>
<tr>
<td>Tp853</td>
<td>218</td>
<td>218</td>
<td>58</td>
<td>26.6</td>
</tr>
<tr>
<td>Average</td>
<td>361</td>
<td>331</td>
<td>64</td>
<td>18.3</td>
</tr>
</tbody>
</table>

N<sub>given</sub> is the total number of provided residue contacts; N<sub>native</sub> is the total number of provided residue contacts where both contact-involved residues are present in the native structure; N<sub>correct</sub> is the total number of correct contacts observed. Residue contacts are defined to exist when the distance between G<sub>a</sub>−C<sub>b</sub>(G<sub>a</sub>−C<sub>b</sub> for glycines) is within 9 Å; the correct ratio is 100× N<sub>correct</sub> / N<sub>given</sub>. 

N<sub>native</sub> is the total number of provided residue contacts; N<sub>correct</sub> is the total number of correct contacts observed. Residue contacts are defined to exist when the distance between G<sub>a</sub>−C<sub>b</sub>(G<sub>a</sub>−C<sub>b</sub> for glycines) is within 9 Å; the correct ratio is 100× N<sub>correct</sub> / N<sub>given</sub>.
Ts, and Tc targets. The objective energy functions to be optimized include distance-restraint energy terms representing the CASP11-provided predicted residue–residue contacts (Tp targets), highly ambiguous and sparse atom–atom contacts (Ts targets), and correct residue–residue contacts (Tc targets).

This article is organized as follows: In the “Methods” section, we describe details of Tp, Ts, and Tc protocols, focusing on differences among these protocols in their energy functions. In the “Results” and “Discussion” section, the overall results of our approach are analyzed in terms of (1) model accuracies among our Tp, Ts, and Tc models and (2) model accuracies between our models and the best of the other groups’ models. In the section titled “What Went Wrong”, we describe what we learned from our mistakes pertaining to early Ts targets and how we improved our algorithms. Finally, future direction of combining the proposed method and NMR experimental data is discussed to advance the field of hybrid methods for protein structure determination.

**METHODS**

There were three types of contact-assisted protein targets in CASP11, namely Tp, Ts, and Tc targets. Depending on the type of target, we used an appropriately designed energy function. Below, three types of energy functions are described along with the rationale for the construction of the functions. In all three cases, the CSA global optimization method was used to generate protein 3D models through the optimization of the energy function. All three energy functions include $E_{\text{steric}}$, $E_{\text{repul}}$, and $E_{\text{phys}}$. The $E_{\text{steric}}$ term, taken from the MODELLER energy function, corresponds to the stereo-chemical terms related to bond lengths, bond angles, torsion angles, and improper torsion angles according to CHARMM22. $E_{\text{repul}}$ corresponds to the repulsive part of the van der Waals potential. It should be noted that these two terms are included in all of our 3D modeling procedure for all the targets.

Our server protocol nns and human protocol LEE followed identical procedures. The only difference between them was the running time of the computation. Typically, two to three days of computation were spent for nns models and up to 7 days for LEE models. For early targets, less than 1 day of computation was performed for nns models mainly due to the incompleteness of the procedure. For many targets, especially earlier ones, LEER models are identical to LEE models. For LEER modeling, LEE models were taken as input models and we then followed the identical procedure as discussed in our other CASP11 papers that was utilized for the regular human target. Since the backbone quality of LEE and LEER models are quite similar to each other, in this work we focus mainly on the procedure and the results of nns and LEE.

For the evaluation of the protein 3D models, in most cases, we used TM-scores that was downloaded from http://zhanglab.ccmb.med.umich.edu/TM-score. The TM-score program aligns atoms of two given protein structures in a sequence-dependent manner and produces a score in the range of 0–1, where 1 corresponds to a perfect match between two structures. One advantage of TM-score is that it is designed to be independent of the protein size. TM-score of 0.7 means that about 70% of $C_a$ atoms are aligned regardless of the protein size. On the other hand, for a similar case, RMSD can be small/large for small/large proteins. In addition to TM-score, GDT-TS, IDDT, and RMSD were also used. Since no standard programs for GDT-TS and IDDT are available, their scores were downloaded from the CASP11 homepage. GDT-TS is the major CASP assessment measure for 3D model evaluation and is a global-distance test score measured by the percentage of aligned $C_a$ atoms using four distance cutoff values of 1, 2, 4, and 8 Å. The local distance difference test, IDDT, is a superposition-free score for comparing protein structures by evaluating local distance differences of all atoms.

**Tp protocol**

For Tp targets, predicted residue–residue contact information was provided. We used the following energy function to incorporate the information:

$$E_{\text{Tp}} = E_{\text{steric}} + E_{\text{repul}} + E_{\text{contact}} + E_{\text{phys}}$$

Here, $E_{\text{contact}}$ is defined as

$$E_{\text{contact}} = w_{\text{contact}} \sum_i \left\{ \begin{array}{ll} (r_i - r_{\text{upper}})^2 & r_i > r_{\text{upper}} \\ (r_i - r_{\text{upper}})^2 + \sigma^2 & \text{otherwise} \end{array} \right.$$  

where $r_i$ is the distance between two $C_a$ ($C_a$ for glycine) atoms for each contact $i$; $r_{\text{upper}}$ is the upper bound ($r_{\text{upper}} = 4.5$ Å) of the given contact with $\sigma = 4.0$ Å. The weight factor $w_{\text{contact}}$ was set to 16.0 kcal mol$^{-1}$. It should be noted that many of the provided residue–residue contacts were likely to be incorrect (see Table I), and we did not want to penalize contact violation too excessively. For this purpose, the functional form of the Lorentzian function$^{22}$ was used for $E_{\text{contact}}$.

$E_{\text{phys}}$ includes all the necessary modeling energy terms for free modeling (ab initio modeling) including: dynamic fragment assembly (dfa) terms,$^{19}$ the dfire statistical potential term,$^{33}$ the hydrogen bonding term,$^{34}$ and the GOAP term$^{35}$ as follows:

$$E_{\text{phys}} = E_{\text{dfa}} + E_{\text{dfire}} + E_{\text{GOAP}} + E_{\text{bond}}$$
methods.\textsuperscript{22,28,29} By applying CSA, we searched for protein 3D models which were diverse and had low energies according to the energy function of Eq. (1).

For initial structures, we used fold recognition models generated by our \textit{nms} protocol and all the server models provided from the CASP11 homepage. Using CSA, typically we generated up to 200 structures, which were clustered by the structural similarity. Each cluster was ranked based on the average value of \( E_{\text{contact}} \) and the average QA3 score.\textsuperscript{22} The best cluster was identified and the final models were selected based on \( E_{\text{contact}} \) and the QA3 score in the cluster.

### Table II

**Summary of Structural Qualities of LEE Model1 is shown for 19 Ts Targets**

<table>
<thead>
<tr>
<th>Target ID</th>
<th>( N_{\text{res}} )</th>
<th>( N_{\text{peak}}/N_{\text{res}} )</th>
<th>( \langle N_{\text{pair}} \rangle )</th>
<th>TM-score</th>
<th>RMSD (Å)</th>
<th>NOE (Å)</th>
<th>Favored (%)</th>
<th>Outlier (%)</th>
<th>Clash</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ts761\textsuperscript{a}</td>
<td>237</td>
<td>13.1</td>
<td>9.4</td>
<td>2.1</td>
<td>0.9084</td>
<td>2.1</td>
<td>0.000</td>
<td>83.8</td>
<td>1.0</td>
</tr>
<tr>
<td>Ts763</td>
<td>130</td>
<td>15.6</td>
<td>5.7</td>
<td>0.8589</td>
<td>2.2</td>
<td>0.000</td>
<td>89.6</td>
<td>1.6</td>
<td>0.00</td>
</tr>
<tr>
<td>Ts777\textsuperscript{a}</td>
<td>345</td>
<td>7.0</td>
<td>18.1</td>
<td>0.7113</td>
<td>5.6</td>
<td>0.000</td>
<td>73.3</td>
<td>9.0</td>
<td>0.00</td>
</tr>
<tr>
<td>Ts785</td>
<td>112</td>
<td>6.2</td>
<td>5.8</td>
<td>0.8290</td>
<td>4.0</td>
<td>0.001</td>
<td>94.6</td>
<td>1.8</td>
<td>0.00</td>
</tr>
<tr>
<td>Ts794</td>
<td>462</td>
<td>6.8</td>
<td>26.7</td>
<td>0.6290</td>
<td>4.4</td>
<td>0.002</td>
<td>86.5</td>
<td>3.3</td>
<td>0.28</td>
</tr>
<tr>
<td>Ts800</td>
<td>212</td>
<td>6.9</td>
<td>13.5</td>
<td>0.9334</td>
<td>1.6</td>
<td>0.000</td>
<td>93.3</td>
<td>0.5</td>
<td>0.00</td>
</tr>
<tr>
<td>Ts802\textsuperscript{b}</td>
<td>118</td>
<td>4.5</td>
<td>3.8</td>
<td>0.7815</td>
<td>2.7</td>
<td>0.000</td>
<td>89.7</td>
<td>3.5</td>
<td>0.00</td>
</tr>
<tr>
<td>Ts804\textsuperscript{b}</td>
<td>194</td>
<td>4.6</td>
<td>8.8</td>
<td>0.7350</td>
<td>4.9</td>
<td>0.000</td>
<td>89.1</td>
<td>4.2</td>
<td>0.00</td>
</tr>
<tr>
<td>Ts810</td>
<td>113</td>
<td>6.5</td>
<td>4.9</td>
<td>0.7478</td>
<td>5.2</td>
<td>0.000</td>
<td>92.8</td>
<td>0.0</td>
<td>0.00</td>
</tr>
<tr>
<td>Ts814</td>
<td>397</td>
<td>5.8</td>
<td>17.8</td>
<td>0.9125</td>
<td>2.6</td>
<td>0.000</td>
<td>86.8</td>
<td>5.3</td>
<td>0.00</td>
</tr>
<tr>
<td>Ts818</td>
<td>134</td>
<td>3.9</td>
<td>4.3</td>
<td>0.8676</td>
<td>2.1</td>
<td>0.000</td>
<td>92.4</td>
<td>0.8</td>
<td>0.00</td>
</tr>
<tr>
<td>Ts824</td>
<td>108</td>
<td>4.8</td>
<td>3.1</td>
<td>0.8332</td>
<td>2.3</td>
<td>0.000</td>
<td>94.3</td>
<td>0.9</td>
<td>0.00</td>
</tr>
<tr>
<td>Ts826\textsuperscript{b}</td>
<td>201</td>
<td>8.3</td>
<td>14.4</td>
<td>0.7722</td>
<td>4.5</td>
<td>0.002</td>
<td>87.4</td>
<td>1.0</td>
<td>0.00</td>
</tr>
<tr>
<td>Ts827</td>
<td>158</td>
<td>6.9</td>
<td>7.5</td>
<td>0.4627</td>
<td>8.9</td>
<td>0.003</td>
<td>80.8</td>
<td>3.2</td>
<td>0.00</td>
</tr>
<tr>
<td>Ts825</td>
<td>404</td>
<td>8.7</td>
<td>21.9</td>
<td>0.5098</td>
<td>1.9</td>
<td>0.004</td>
<td>90.3</td>
<td>2.7</td>
<td>0.62</td>
</tr>
<tr>
<td>Ts806</td>
<td>256</td>
<td>7.0</td>
<td>14.9</td>
<td>0.8905</td>
<td>2.4</td>
<td>0.001</td>
<td>90.6</td>
<td>1.6</td>
<td>0.00</td>
</tr>
<tr>
<td>Ts832</td>
<td>209</td>
<td>7.0</td>
<td>12.0</td>
<td>0.8410</td>
<td>2.5</td>
<td>0.002</td>
<td>87.9</td>
<td>2.4</td>
<td>0.00</td>
</tr>
<tr>
<td>Ts767</td>
<td>274</td>
<td>5.7</td>
<td>9.4</td>
<td>0.8958</td>
<td>2.3</td>
<td>0.002</td>
<td>87.9</td>
<td>1.5</td>
<td>0.45</td>
</tr>
<tr>
<td>Ts812</td>
<td>183</td>
<td>5.4</td>
<td>5.6</td>
<td>0.8003</td>
<td>3.4</td>
<td>0.001</td>
<td>82.3</td>
<td>5.0</td>
<td>0.00</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>224</td>
<td>7.1</td>
<td>10.9</td>
<td>0.8133</td>
<td>3.6</td>
<td>0.00</td>
<td>89.3</td>
<td>2.5</td>
<td>0.06</td>
</tr>
</tbody>
</table>

\textsuperscript{a}Targets are listed according to their release dates. \( N_{\text{res}} \): number of residues; \( N_{\text{peak}}/N_{\text{res}} \): number of peaks per residue; \( \langle N_{\text{pair}} \rangle \): average number of ambiguous atom–atom pairs for each NOE peak; RMSD (Å): \( C_{\alpha} \) RMSD between model and native structure; NOE (Å): RMSD of NOE violation; favored (%): the percentage of residues in the favored region of the \( \phi/\psi \) Ramachandran diagram; outlier (%): the percentage of residues in the outlier region of the \( \phi/\psi \) Ramachandran diagram; Clash: number of atom–atom steric clashes per 1000 atoms. \textsuperscript{b}Model is recalculated using the latest Ts protocol without the erroneous screening, and the numbers in parenthesis are from the submitted model that used the erroneous screening. \textsuperscript{c}Native structure is not disclosed yet and is substituted by LEE model1 of Tc target to calculate TM-score and RMSD.

The modeling of Ts targets is closely related to the NMR structure determination using NOE distance restraints. In the usual NMR structure determination situation using NOE distance restraints, measured NOE peaks are all unambiguously assigned to their atom–atom pairs in a one-to-one fashion. However, the contact restraints provided for Ts targets were highly ambiguous, meaning that many atom–atom pairs were assigned to a peak. As shown in Table II, the average number of atom–atom pairs per NOE peak is about 10.9.

For the Ts modeling, we employed a method based on our NMR structure determination,\textsuperscript{14} where the CSA optimization is applied to the following energy function:

\[
E_{\text{Ts}} = E_{\text{stereo-chemistry}} + E_{\text{vdw}} + E_{\text{chiral}} + E_{\text{CMAP}} + E_{\text{NOE}}
\]

where \( E_{\text{stereo-chemistry}} \) constrains bond lengths, bond angles, torsion angles and improper torsion angles, and is taken from the CHARMM22 force field. \( E_{\text{vdw}} \) is the repulsive part of the van der Waals potential. The third term \( E_{\text{chiral}} \) is to keep the chirality of the amino acid residue in the \( L \) form.\textsuperscript{14} Since the driving force for the structure modeling of Ts targets is hydrogen–hydrogen distance restraints, without \( E_{\text{chiral}} \) the mirror image of a given structure (either in \( L \) or \( D \) form) will be as stable as the given structure.

The next term, \( E_{\text{CMAP}} \) was used to provide a cross-term correction for two adjacent torsion angles,\textsuperscript{36} which is a standard stereo-chemical term in the CHARMM force field. Finally, for a given set of ambiguous restraints, the most critical term, \( E_{\text{NOE}} \), is defined as follows:

\[
E_{\text{NOE}} = \sum_i \begin{cases} 
  k_i (R_i - d_i^\text{upper})^2, & R_i > d_i^\text{upper} \\
  0, & d_i^\text{lower} < R_i < d_i^\text{upper} \\
  k_i (R_i - d_i^\text{lower})^2, & R_i < d_i^\text{lower}, 
\end{cases}
\]

where the sum is over all the Ts restraints and the index \( i \) represents each restraint. The effective distance \( R_i \) is calculated by \( R_i = (\sum_j d_j^{-6})^{-1/6} \), where the sum over \( j \) is...
taken for all possible distance pairs provided for the $i$th restraint. For each restraint $i$, $d_{i}^{\text{upper}}$ and $d_{i}^{\text{lower}}$ respectively correspond to the upper and the lower bounds of the pairwise atom distance. Here, $k_i$ is set to $10^3/d_i^{\text{upper}}\text{kcal mol}^{-1}\AA^{-1}$. We note that the conventional usage of the effective distance $R_i$ suits our purpose of solving the two-tier optimization problem, where the variation of an assigned peak to an atom–atom pair is possible. Depending on the initial protein structure, the minimization of Eq. (4) will provide various peak assignment possibilities (i.e., the identity of the atom–atom pair contributing the most in $d_i^{2,6}$ varies).

Low-energy protein structures were searched again by applying CSA to Eq. (4). For initial structures, we used the same initial models as used in the Tp protocol. We generated up to 700 structures which were subsequently reminimized using the LBFGS local minimization method on the new energy function: $E_{\text{TP}} = E_{\text{TS}} + E_{\text{electro-static}} + E_{\text{gbsa}}$ where the second and the third terms respectively correspond to the CHARMM22 electro-static potential and gbsa solvation energy. Final models were selected based on the values of $E_{\text{TS}}$ and the Ramachandran-favored score calculated by the MolProbity program.

**Tc protocol**

CASP11 organizers provided 24 Tc targets with correct residue–residue contact information. Among them, 17 were released after their corresponding Tp targets were released. This means that for these 17 Tc targets, the Ts information was available at the time of the Tc prediction, for which we used the energy function of $E_{\text{Tc}} = E_{\text{TS}} + E_{\text{contact}}$ and followed the identical Ts modeling protocol. Here $E_{\text{contact}}$ is identical to the one in Eq. (1) except that $r_{\text{upper}}$ was set to 7.0 Å and $w_{\text{contact}}$ was set to 32.0 kcal mol$^{-1}$. For initial structures, we also used the same initial models as used in the Tp protocol.

For the other 7 Tc targets, during the period of the Tc prediction, no Ts information was available. For them, we followed the Tp modeling protocol using $E_{\text{Tp}}$ in Eq. (1), again using $r_{\text{upper}} = 7.0$ Å and $w_{\text{contact}} = 32.0$ kcal mol$^{-1}$.

**Figure 1**

A summary of the modeling results of our nns model1 (a) and best model (b) is shown for all 25 regular (T0) targets, 24 Tp targets, 19 Ts targets, and 24 Tc targets in terms of their TM-scores. For 20 out of 25 T0 targets, the TM-score of nns model1 is less than 0.4 meaning that relevant templates of these 20 targets were either missed or absent in PDB. On average, Tp models are slightly better than T0, but Ts or Tc models are greatly improved over T0 targets. For most targets, target release sequence was Tp, Ts, and Tc. But for Targets marked by asterisk (Ts767 and Ts812) it was Tp, Tc and Ts, and their Ts models are more accurate than corresponding Tc models.

**Figure 2**

The energy landscape of Tp767 is shown. For this particular target, its second domain, that is, Tp767-D2 was particularly well predicted in nns model2, and we show the TM-score of Tp767-D2 in terms of the energy of the whole chain. The nns model2 (red) of Tp767-D2 is shown in the cartoon representation along with its superimposed native structure (green). The energy of the native structure is $-77925$, which is off from the energy scale. When we minimized the native structure using our energy function of Eq. (1), the TM-score was slightly lowered to the value of 0.9887, and the energy became $-85045$ (shown by an arrow in the figure). All energy values are in the unit of kcal/mol.
RESULTS AND DISCUSSION

Tp targets were released after their corresponding T0 targets expired. However, as will be discussed later, only a small fraction of the provided predicted contact information was correct in most cases (see Table I below). After Tp targets expired, Ts and/or Tc targets were released. For T0767 and T0812, Tc targets were released first before their Ts targets were released, and for most of the other targets, the Ts target release was followed by the Tc target release. After Ts/Tc targets expired, corresponding Tc/Ts targets were released. We note that some contact-assisted targets (Tp/Ts/Tc) were not released for various scheduling reasons (see Fig. 1).

Figure 1 shows the nns server prediction results for all contact-assisted targets as well as the corresponding regular T0 targets in terms of TM-score. We observed that, on average, the accuracies of Tp targets (shown in circles) were slightly higher than those of T0 targets (shown in squares). In contrast, we found that the accuracies of Ts models (shown in diamonds) which utilized ambiguous NOE restraints and those of the Tc models (shown in triangles) which utilized correct contact restraints were greatly improved. The improvements of Ts761 and Ts777 were less significant than other Ts targets; this is due to an erroneous procedure we took in the early stage of the Ts prediction period. The procedure mistakenly excluded meaningful restraints during the CSA optimization process. This is further discussed in the next section, “What Went Wrong”.

As stated in the previous section, for the Tc targets, 17 were released after Ts targets, whereas 7 (Tc767, Tc812, Tc831, Tc834, Tc836, Tc848, and Tc853) were released alone or before the release of Ts targets. For Tc or Ts models that utilized both of the Ts and Tc information, the models were more accurately predicted compared to those utilizing only one (Ts or Tc) information.

Figure 2 shows the structural superposition of Tp767-D2 (nns model2 in red) with the native structure (in green) together with the energy landscape of the final
100 CSA models. We used the initial structures obtained from 88-fold recognition models (from the nns protocol) and 100 CASP11 server models. TM-scores of these models ranged from 0.10 to 0.20. The TM-score of the best final model was 0.4237, and the core regions in particular with the beta-strands around the helices were well predicted. The energy landscape of this target shows good correlation between the total energy and the TM-score. We also note that, for this particular target, the model with the lowest energy was the best in TM-score, but we unfortunately chose this best model as model2 by using other selection measures (see the section of “Tp Protocol”).

Table II shows the six structural qualities of the LEE model1 for 19 Ts targets. The qualities are: the model’s TM-score, Cα RMSD from its native structure (in Å), average RMSD of NOE violation from the simulated NOE data (in Å), percentage of residues in the favored φ/ψ Ramachandran region, percentage of residues in the outlier φ/ψ Ramachandran region, and number of atom–atom steric clashes per 1000 atoms. Our models for Ts761 and Ts777 were significantly poor in the above qualities (see Fig. 1 and Table II) because they were built using erroneous procedure. New models generated after fixing the error were much improved.

Since the native structures of three Ts targets Ts802, Ts804, and Ts826 (out of 19) had not been and still are not available as of this writing, their structural qualities were measured in comparison to their corresponding Tc models. For the 19 Ts targets, the average number of NOE peaks per residue (N_{\text{peak}}/N_{\text{res}} in Table II) and the average ambiguity (\langle N_{\text{pair}} \rangle in Table II) were 7.1 and 10.9, respectively, which demonstrates the sparsity and ambiguity of the restraints provided. In comparison, N_{\text{peak}}/N_{\text{res}} and \langle N_{\text{pair}} \rangle from the NOE data (used for a typical NMR structure in PDB) are about 15 and 1, respectively. This demonstrates that the data used to determine a typical NMR structure in PDB contains more peaks per residue with far less ambiguity. The average values of TM-score and RMSD are respectively 0.8133 and 3.6 Å. The NOE values in Table II indicate the RMSD of the violated distances. The violated distance is zero if the actual distance falls between the lower and upper bounds. The average NOE violation is only 0.001 Å which means that our models satisfy essentially all of the given restraints. Our models exhibit, on average, a high Ramachandran favored portion (89.34), low outlier portion (2.52), and extremely low clash score (0.06). It should be noted that we have not used any energy terms that directly control the Ramachandran portion of the favored and outlier regions. Therefore, the improvements in the Ramachandran portion must be an effect of the properly combined various energy terms used in our model building procedure.

Figure 3 shows stereo views of six Ts models in their cartoon representations together with their native structures in superposition. TM-score and RMSD are also shown. Ts794 is a large two-domain protein with 462 residues, and its Cα RMSD value is only 4.4 Å. The domain in the right upper part of Ts794 shows less secondary structure formation, for which additional refinement procedure is necessary for improvement. The structure of another large protein, target Ts835, was determined with the accuracy of less than 2.0 Å of Cα RMSD. Figure 4 shows the history of the lowest energy model for Ts767. After about one day of the CSA optimization using 210 CPU cores of Intel Xeon X5670 at 2.9 GHz were used for the calculation. The cartoon representation of the lowest energy model after 3, 10, 20, 24 h, and 6 days is shown. The percentage of residues in the favored region of the φ/ψ Ramachandran diagram is also indicated.
these Tc models were generated using only the Tc information (correct residue–residue contact information), and for these particular two, their Ts targets were not tested in CASP11. Considering the relatively small number of provided residue–residue contacts for Tc834 and Tc836, their TM-scores of LEE model1s (0.8106 for Tc834 and 0.7011 for Tc836) are quite high. For target Tc834, Figure 7 shows the energy landscape of 200 final CSA models, from which we observed a good correlation between the energy and TM-score.

**WHAT WENT WRONG**

Ts761 and Ts777 are two large targets (with 237 and 345 residues, respectively) that were released early during CASP11, and we failed to model them properly (see Fig. 1). The top panel of Figure 8 shows our submitted model of Ts761. Its Cα RMSD over the native structure is >17 Å and its Ramachandran favored portion is only 63%. This poor modeling was due to the erroneous filtering process used in our modeling protocol of the two Ts targets. To reduce the ambiguity level of provided restraints and to speed up the calculation, we intended to filter out incorrect atom pairs by comparing the model structures generated in the early stages of CSA with the given restraints. For Ts761, we were provided with 3106 NOE peaks ambiguously assigned to 112,130 atom pairs. We filtered out 58771 atom pairs among 1971 peaks, then-believing confidently that these pairs were wrong since they were inconsistent with all of the 100 interim CSA models with the additional 2 Å allowance. However, when checked after the native structure was released, about 170 (0.3% of 58771) were found to
be wrongly eliminated by the filtering process, causing other wrong atom pairs to be assigned to those 170 peaks. This resulted in the wrong modeling irreversibly.

Although the NOE violation (measured by the ambiguous NOE restraints) of the submitted model was as low as 0.004 Å, the model was of poor secondary structures, relatively low Ramachandran favored portion (65%), and high outlier portion (8%). In Figure 8, the model on bottom shows the Ts761 model recalculated after CASP11 without the erroneous filtering process. The recalculated model is of 2.1 Å of RMSD with rich secondary structures including a beta-wall like structure in the middle and α-helices.

CONCLUSIONS

We have applied the global optimization method of conformational space annealing (CSA) to the energy functions appropriately designed to the modeling of contact-assisted targets in CASP11. The protein model accuracy depended on the quality of the given contact information. For all the Tp targets that were provided with predicted contact information, physical energy terms developed for free modeling were used together with their contact energy terms that do not excessively penalize the violation of given contacts. This was to deal
with potentially wrong contact information, as it turned out that only about 18% of the provided predicted contact information was correct. On average, Tp models were improved over T0 models, but the improvement was significant only for a few targets.

In order to deal with the ambiguous NOE data of Ts targets, we transformed the modeling of a Ts target into a two-tier optimization problem which, at first glance, appeared to be an impossible task to solve by direct method. However, using only the provided data, we managed to solve the two-tiered problem where the double complexity of the peak assignment as well as the 3D arrangement of the protein’s atoms were treated simultaneously. The key energy term for shaping the protein into its native structure was the NOE violation term which was calculated in terms of the effective distance \( R = \left( \sum_j d_{ij}^{-6} \right)^{-1/6} \), where the sum is taken for all ambiguous distance pairs provided. Even with sparse and very ambiguous restraint information, the CSA method was quite successful in finding 3D models consistent with the given restraints. We note that attempts to solve the protein structure modeling utilizing sparse and ambiguous NMR data by direct method had never been tried before. It was encouraging to find out that rather accurate 3D protein models were generated, typically with RMSD values between 1.6 and 3.4 Å for most Ts targets (except the first two large targets released in the early stage that we covered in “What Went Wrong”). An attempt to eliminate seemingly wrong atom–atom pairs in the early stages of CSA optimization turned out to be rather problematic because any elimination of true atom–atom pairs inevitably introduces as many wrong pairs; this ruins the whole modeling procedure. It would be interesting to find out (1) how well our proposed method would work for protein structure determination using real NMR data and (2) how much additional improvement can be achieved by combining our new method with the evolutionary residue–residue coupling method.

For Tc targets, further-refined structures over Ts models were obtained by using the additional correct contact information provided. Based on the proposed methods for the contact-assisted category, we can conclude that both efficient global optimization and a utilization of proper energy functions (that describe the given experimental information) are critical for generating accurate models.

**ACKNOWLEDGMENTS**

We thank Andrew Brooks for his critical reading of and suggestions for the manuscript. We thank KIAS Center for Advanced Computation for providing computing resources.

**REFERENCES**


