Supplementary Material for

"Constructing Structure Ensembles of Intrinsically Disordered Proteins from Chemical Shift Data"

Huichao Gong[1]
Sai Zhang[1]
Jiangdian Wang[2]
Haipeng Gong[3,4]
Jianyang Zeng[1,4,*]

January 14, 2015

The following is supplementary material which provides additional information to substantiate the claims of the paper. Section 1 presents the procedure of validating our least squares method using alpha-synuclein through reference ensemble. Section 2 describes the details of validating our elastic net method using alpha-synuclein through reference ensemble.

1 Validating the least squares method through Reference ensemble

We ran a 30ns MD simulation for alpha-synuclein and saved the coordinates every 3 picoseconds, which yielded 10,000 structures as our initial structure pool. To ensure the constraints derived from experimental constraints outnumber the weights to be estimated, we set the cutoff parameter to be 6.0 Å. Then 71 structures were selected using the structure clustering procedure as described in Section 2.3. Next, we used these 71 structures as the "true" conformations in the reference ensemble, and set their "true" weights. To investigate the impact of different settings of "true" weights, we tested two typical distributions, including uniform and Gaussian distributions.

After constructing the reference ensemble, the "experimental" data were synthesized. We first predicted the chemical shifts of backbone atoms, including HN, HA, CA and N, using the chemical shift prediction software SHIFTX2 [1] for each conformation in the reference ensemble. Next, we synthesized the chemical shift of a backbone atom in the i^{th} residue, denoted by b_i^T, based on the following equation:

$$b_i^T = \sum_{j=1}^{n} a_{ij} w_j^T + N(0, \sigma^2), \tag{1}$$

where n stands for the total number of conformations in the ensemble, a_{ij} represents the SHIFTX2-predicted chemical shift of the corresponding atom in the i^{th} residue of the j^{th} structure in the reference ensemble,

[1]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, 100084, P. R. China
[2]Biostatistics and Research Decision Sciences — Asia Pacific, Merck Research Laboratory, Beijing, 100015, China
[3]School of Life Sciences, Tsinghua University, Beijing, 100084, China
[4]MOE Key Laboratory of Bioinformatics, Tsinghua University, Beijing, 100084, China
*Corresponding authors: Jianyang Zeng, zengjy321@tsinghua.edu.cn.

and $N(0, \sigma^2)$ stands for Gaussian noise with standard deviation σ which simulates experimental noise. To consider the impact of different levels of noise, we set σ from 0.01 to 0.06 with step size 0.01. Here, we only presented the results of three specific values: 0.01, 0.02 and 0.03, since they were enough to depict the trend.

We sampled b_i^T for 20 times and back-computed the corresponding optimal weights using our methods described in Section 2.5. Then, these 20 sets of weights were averaged as the output weights $\boldsymbol{w^C}$. After this, as in [2], we used the Jensen-Shannon divergence to evaluate the difference between the computed and reference ensembles:

$$\Omega^2\left(\boldsymbol{w^C}, \boldsymbol{w^T}\right) = S\left(\frac{\boldsymbol{w^C} + \boldsymbol{w^T}}{2}\right) - 0.5S\left(\boldsymbol{w^C}\right) - 0.5S\left(\boldsymbol{w^T}\right), \tag{2}$$

where $\boldsymbol{w^T}$ represents the "true" weight vector of the selected structures, and $S(\boldsymbol{w}) = -\sum_{j=1}^{n} w_j \log(w_j)$, in which w_j means the j^{th} element of vector $\boldsymbol{w}$.

In addition, the back-computed chemical shift of a backbone atom of the i^{th} residue, denoted by b_i^C, was computed by:

$$b_i^C = \sum_{j=1}^{n} a_{ij} w_j^C. \tag{3}$$

Furthermore, the RMSD between back-computed and "true" chemical shifts was computed as follows:

$$RMSD = \sqrt{\frac{1}{m}\left(\boldsymbol{b^C} - \boldsymbol{b^T}\right)^2}, \tag{4}$$

where m is the total number of elements in vector $\boldsymbol{b^C}$ or $\boldsymbol{b^T}$.

To ensure that our results were independent of the initial choices of $\boldsymbol{w^T}$, each test was repeated for 50 times with different random seeds. We first evaluated the performance of the least squares method and compared to that of the Monte Carlo approach. As shown in Fig.2, our least squares method significantly outperformed the Monte Carlo approach in all different settings, and the weights computed by our algorithm agreed better with the "true" weights which were generated following both uniform and Gaussian distributions. When the standard deviation of Gaussian noise in synthesized experimental data increased, the performance of our algorithm was gradually degraded towards the Monte Carlo approach. However, the computed weights were still in an acceptable accuracies, even for the largest Gaussian standard deviation (0.03ppm). These results indicated that our algorithm can be used to reliably compute the accurate weights of individual conformations in an IDP ensemble.

2 Validating the elastic net method through Reference ensemble

We ran a 30ns MD simulation for the K18 domain of Tau protein and saved the coordinates every 3 picoseconds, which generally yielded 10,000 structures as our initial structure pool. Then we selected 882 structures using the structure clustering procedure described in Section 2.3, with the cutoff parameter set to be 2.1 Å. Next, we used these 882 structures as the "true" conformations in the reference ensemble, and set their "true" weights. In particular, to focus on the sparsity condition, we set only 10 non-zero weights with Gaussian distribution for these structures, and normalized them to sum up to 1. In the elastic net algorithm, the α was chosen to be 0.95 via a cross-validation procedure.

The Jensen-Shannon divergence between computed and "true" weights and the RMSD between back-computed and "true" experimental data are shown in Fig.4. We found that the elastic net approach significantly outperformed the Monte Carlo approach in this sparsity condition.

Suplementary References

[1] Beomsoo Han *et al. Journal of Biomolecular NMR*, 50(1):43–57, 2011.

[2] Charles K. Fisher *et al. Journal of the American Chemical Society*, 132(42):14919–14927, 2010.