BMI 206 Final Project Report

Seth D. Axen

Fall, 2014

Introduction

The Nup84 heptameric complexes form the outer ring of the nuclear pore complex. Electron micrographs (EMs) of two conformations of the Nup84 complex were solved in 2009[1], and a subsequent study used the Integrative Modeling Platform (IMP)[2] to predict a low-resolution structure of the complex based on domain mapping, 2D EM, and crystal structures[3]. In the assigned paper, a higher resolution structure of the Nup84 complex was predicted by additionally using two types of chemical cross-links, resulting in 16,000 structures, 5,000 of which were used to generate a final ensemble solution.[4]

Methods

Running IMP

First, the original paper's GitHub repository was cloned[5]. Then two replicates were run with IMP using all data constraints save for the crystallographic interface constraints, for increased computational efficiency. To enable replica exchange, these jobs were run on the cluster for 24 hours with 4 nodes, resulting in 681 and 317 solutions, respectively. A third replicate was then run for 36 hours with 6 nodes, resulting in 682 solutions. These replicates were subsequently analyzed separately as well as together.

Clustering

For each IMP run, the top 250 best scoring solutions were clustered with k-means into a single cluster, using root mean squared deviation (RMSD) as a distance metric. In addition, the top 100 solutions were divided by k-means into two clusters. The top 100 merged replicates results were also structurally aligned by RMSD prior to clustering by k-means.

Localization Density

Localization densities were calculated using IMP's built-in functionality and visualized in Chimera[6] (Fig. 1).

Precision Analysis

For each of the three replicates as well as the merged run, the pairwise inter-cluster and intra-cluster precision (Fig. 2) were calculated as the average pairwise distance root mean squared deviation (dRMSD) between all solutions in the two clusters or in the single cluster, respectively.

10-fold Cross-validation of cross-links

Each of the two types of cross-links were divided into 10 random, non-overlapping groups. 10 training and testing sets of cross-links were consequently generated, where each training set consisted of 90% of the cross-links for each type, and the test set consisted of the remaining 10%. IMP was run using each of these 10 training sets as constraints, and the results were clustered as with the three replicates. The percent of the

remaining 10% of cross-links that were satisfied for each run was calculated, and these 10 percentages were averaged to generate a cross-validation error.

Data deposition
All data, analysis, and scripts were deposited in GitHub.[7]

Results
Qualitative comparison of results
The three replicate IMP runs generated 681, 317, and 682 Nup84 complex solutions, respectively. While the localization density from the first four solutions from Run 1 appeared qualitatively similar to the results from the paper (Fig. 1A, 1E), this similarity decreased when the 250 best solutions from Run 1 (Fig. 1B) or the 250 best solutions from all replicate runs (Fig. 1C) were used. To explore whether the heterogeneity across the three runs was caused by misaligned solutions, the 100 best solutions were aligned by RMSD. The resulting ensemble exhibited a localization density vastly more similar to that found in the original paper than the solutions did before the alignment (Fig. 1D).

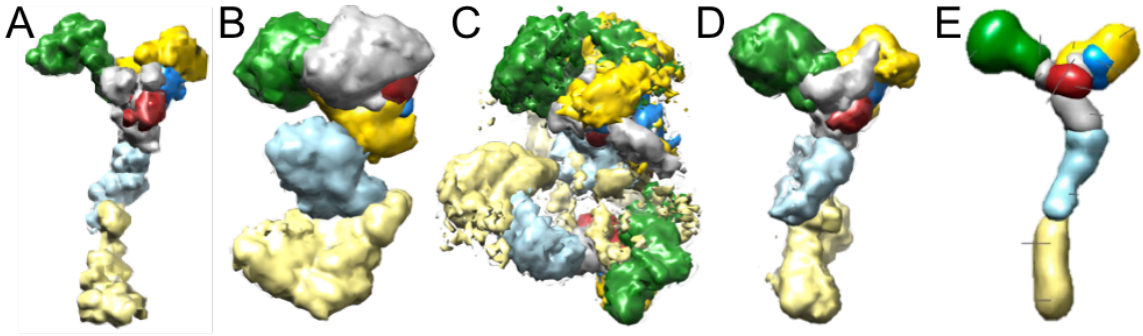

Fig. 1. Nup84 solution localization densities from A) the four best solutions from run 1, B) the 250 best solutions from run 1, C) the 250 best solutions from all runs, D) the 100 best solutions from all runs, aligned by RMSD, E) the original publication.

Analysis of precision
The precision, the average pairwise dRMSD between two sets of solutions, represents an estimate of the structural heterogeneity between the two sets. When comparing a cluster with itself, this quantity represents the self-consistency of the solutions in the clusters with the other solutions in the cluster. Inter- and intra-cluster precision values were calculated for the whole Nup84 complexes and the central hub (all subunits but Nup133 and Nup84), comparing each of the clusters generated using k-means at a cut-off of two. All clusters have relatively low intra-cluster dRMSD values (Fig. 2A). However, in the individual runs, the clusters also have relatively low dRMSD values with respect to one another. When comparing the individual run clusters with one another, all have very high dRMSD values, indicating low precision (Fig. 2A). While these trends continue when precision is calculated based on the hub, they are less pronounced (Fig. 2B).

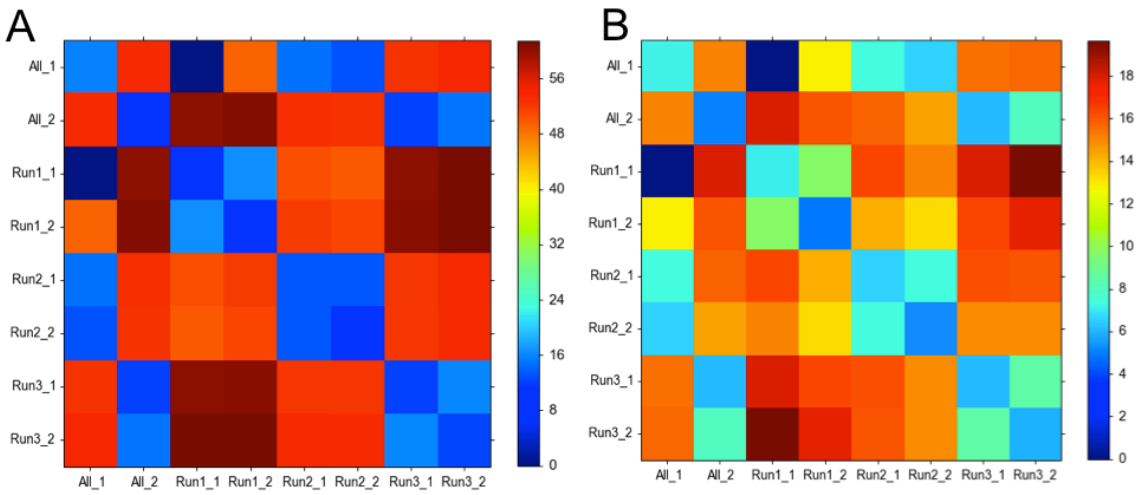

Fig. 2. Intra- and inter-cluster precisions for A) the whole Nup84 complex and B) the Nup84 hub. Each row-column pair indicates a set of structures for which the dRMSD was calculated. A low dRMSD (blue) indicates high precision, and a high dRMSD (red) indicates low precision.

Cross-validation of cross-links

10-fold cross-validation using cross-links yielded a cross-validation error of 3.8%.

Discussion

Overall, the above analysis successfully reproduced the overall trends observed in the original paper. While this analysis did not converge upon a solution, this is likely because the original publication generated nearly ten times as many solutions. Furthermore, while the scripts provided in the corresponding GitHub repository did not contain any commands for RMSD alignment of solutions, such an alignment greatly improved the above results, and the author speculates that the original analysis did include such an alignment step.

That insufficient solutions have been generated and/or that k-means clustering without first aligning with RMSD does not result in reproducible clusters is apparent from Fig. 2, as one would expect a one-to-one relationship between a pair of clusters from one run with a pair of clusters from another run if the clustering was successfully generating similar ensembles of solutions in each run. RMSD alignment could be used to generate more reliable clusters, and if this lack of one-to-one correspondence continues, this would confirm that the space of solutions given the restraints has not yet been sufficiently explored.

Cross-validation of cross-links demonstrates that the number of cross-links used as restraints in the analysis in the paper is sufficient to converge upon a solution that satisfies the maximum number of cross-link restraints, as removal of 10% of restraints nevertheless resulted in only 3.8% of those restraints still being met by the final ensemble of solutions. As the original paper noted in the supplement that 20% of cross-links were not satisfied by the final ensemble, this error seems acceptable.

One future direction to enhance this analysis would be to use both solved EM maps of the Nup84 complex by Kampmann and Blobel[1] as additional restraints. In this analysis, I attempted to verify the accuracy of our results against these EM maps but was unable to transform them to a file format compatible with IMP. If this could be achieved in the future, one could finely tune parameters of IMP to generate solutions that better fit these solved structures.

Acknowledgments
I would like to thank Katie Pollard and Andrej Sali for instruction, Ben Webb, SJ Kim, and Elina Tejoe for helpful conversations and technical assistance, as well as Hubert Nethercott for useful discussions.

References

1. Kampmann, M. & Blobel, G. Three-dimensional structure and flexibility of a membrane-coating module of the nuclear pore complex. *Nat. Struct. Mol. Biol.* **16,** 782–8 (2009).
2. Russel, D. *et al.* Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol.* **10,** e1001244 (2012).
3. Fernandez-Martinez, J. *et al.* Structure-function mapping of a heptameric module in the nuclear pore complex. *J. Cell Biol.* **196,** 419–34 (2012).
4. Shi, Y. *et al.* Structural Characterization by Cross-linking Reveals the Detailed Architecture of a Coatomer-related Heptameric Module from the Nuclear Pore Complex. *Mol. Cell. Proteomics* **13,** 2927–43 (2014).
5. Pellarin, R., Tjioe, E. & Kim, S. J. Nup84. *GitHub Repos.* (2014). at <https://github.com/integrativemodeling/nup84>
6. Pettersen, E. F. *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25,** 1605–12 (2004).
7. Axen, S. BMI206 Final Project. *GitHub Repos.* (2014). at <https://github.com/sdaxen/bmi206_final_project>