SUPPLEMENTAL METHODS

Defining edges in the graph for adjacent subunits

Edges in a graph represent adjacency between the vertices. In some contexts it is quite clear whether or not an edge should exist between two vertices (e.g. in a social network you are either friends with someone or you are not). In the biological context of a protein complex there is no universal definition of what constitutes an edge between protein subunits. Even defining the edges between subunits from a known 3D structure of a complex requires a precise threshold on some objective measurement of interaction between pairs of subunits, such as the amount of buried solvent accessible surface area.

In this work we use the following conceptual definition for what constitutes an edge between subunits in a protein complex: An edge exists between a pair of subunits if upon binding they bury a significant amount of solvent accessible surface area (SASA). We qualitatively define 'significant' to mean be an unknown amount such that each Rpt subunit is adjacent to precisely two other Rpt subunits. Since we are interested in inferring topologies of unknown complexes, a precise definition of 'significant' would not be any more useful than a simple qualitative definition; however, here we investigate interface sizes to understand what would be reasonable SASA thresholds for defining adjacency

To estimate the interface sizes between adjacent subunits in the base we performed calculations on the base hexamer subunits in the experimentally determined structure of p97, an abundant hexameric ATPase of the AAA family. Calculations were made on the structure with (PDB code 1E32) {Zhang, 2000} and the sizes of the six interfaces were in the narrow range from 1003 $Å^2$ to 1024 $Å^2$, and 0 $Å^2$ between all nine non-adjacent pairs. Thus, any threshold for defining adjacency between 0 and 1003 $Å^2$ would result in the same graph structure (i.e. cycle of

6). The size of the interface (in $Å^2$) for each pair of subunits (*A*,*B*) was calculated using the solvent accessible surface (SASA) of the monomers and the dimer by: [[SASA(*A*)+SASA(*B*)]-SASA(*AB*)]/2. The SASA calculations were made using the 'measure sasa' command of the Visual Molecular Dynamics (VMD) software {Humphrey, 1996} with default probe radius of 1.4 angstroms.

To present a concrete example of subunit adjacency and the corresponding matrix (which is equivalent to a graph structure) we analyzed the twelve subunit RNA polymerase II (RNAPII) structure (PDB code 1WCM). First, we defined a topology graph of the RNAPII complex based on visual inspection using our rough sense of "significant" subunit interfaces. Next, we sought to identify an objective measure to reproduce our intuitive definition. For this purpose the sizes of the interfaces between all pairs of subunits in the complex were calculated as with the hexameric ATPase in described above, and the complete results are presented in Supplemental Figure 2. There are a total of 66 subunit pairs and 37 have an interface of 0 $Å^2$ (i.e. there is no shared interface) and 29 have an interface > 0 $Å^2$. By far the largest interface, of 9271 $Å^2$, is found between subunits Rbd1and Rbd2, and the smallest non-zero interface, of only 5 $Å^2$, is found between Rbd2 and Rbd8 (Supplemental Figure 2A).

For the sake of objectively calculating a protein topology graph from the 3-D structure of the complex we used a threshold of 250 $Å^2$ to define subunit adjacency (note that any threshold in the range [189,360] $Å^2$ results in the same adjacency matrix – Supplemental Figure 2A-B). This threshold was selected because it is consistent with our definition of the topology graph based on visual inspection of the RNAPII complex. Using this threshold 18 pairs of subunits are considered adjacent and the other 48 are considered non-adjacent. Supplemental Figure 2B presents the resulting adjacency matrix. The adjacency matrix is equivalent to an undirected

graph, as it defines all of the edges in the graph. Supplemental Figure 2C contains the counts of observed cross-links (i.e. a specific lysine pair is only counted once even if it is observed multiple times) between all pairs of subunits in Chen et al. {Chen, 2010}. Using this interpretation of the data there are a total of 65 cross-links. By comparing the adjacency matrix (graph) with the observed cross-linking data we find that 53 of the cross-links come from adjacent pairs of subunits and 12 come from non-adjacent subunits. In the next section we describe how this information is used to establish a confidence interval on the global parameter *p*.

Establishing confidence intervals on the parameter p

In the probabilistic model the value of the global parameter *p* indicates the confidence of the model, where a higher value of *p* means higher confidence; however, the value of *p* is unknown. Thus, we are interested in obtaining point estimates and establishing confidence intervals on the parameter *p*.

Using the subunit adjacency graph of the RNAPII complex described above 53 cross-links come from adjacent pairs and 12 come from non-adjacent pairs. Since *p* is a binomial random variable the probability of observing 53 cross-links (k=53) in 65 random draws (n=65) is only a function of *p*. The probability density function (PDF) on *p* is a beta distribution with α = 54 and ß = 13. The mode of the distribution is 0.82 and the 95% confidence interval on *p* is obtained by integrating over the PDF resulting in the range [0.72,1.0]. The PDF is presented in Supplemental Figure 3A. To obtain this confidence interval we utilized a known protein complex and data from an external group.

To obtain a confidence interval using our own cross-linking data, we make the assumption that the proposed ordering of the hexameric base (Rpt1-2-6-3-4-5) is correct. Under this assumption 10 cross-links come from adjacent pairs and 1 comes from a non-adjacent pair. Thus, the resulting PDF on p is a beta distribution with $\alpha = 11$ and $ß = 2$. The mode of the distribution is $0.\overline{90}$ and the 95% confidence interval on p is obtained by integrating over the PDF resulting in the range [0.66,1.0] as seen in Supplemental Figure 3B.

These confidence intervals on p are used to restrict the search for the maximum likelihood graph of the 19S Rpt base heterohexamer and lid PCI domain-containing heterohexamer to the most likely values of p.

References:

Chen, Z.A., Jawhari, A., Fischer, L., Buchen, C., Tahir, S., Kamenski, T., Rasmussen, M., Lariviere, L., Bukowski-Wills, J.C., Nilges, M., Cramer, P., Rappsilber, J. Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry. EMBO J. 2010 29(4), 717-26.

Humphrey, W.; Dalke, A. and Schulten, K. VMD: Visual molecular dynamics. Journal of Molecular Graphics 1996, 14, 33-38.

Zhang, X., Shaw, A., Bates, P.A., Newman, R.H., Gowen, B., Orlova, E., Gorman, M.A., Kondo, H., Dokurno, P., Lally, J., Leonard, G., Meyer, H., Van Heel, M., Freemont, P.S. Structure of the AAA ATPase p97, Mol.Cell 2000, 6(6), 1473-84.