Inferring Optimally Precise and Maximally Accurate Models from Electron Microscopy Data

by

Charles Harold Greenberg

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biophysics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

[redacted]

Inferring Optimally Precise and Maximally Accurate Models from Electron Microscopy Data

by

Charles Harold Greenberg

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biophysics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

Chair

Committee in Charge

This thesis is dedicated to my family.

Especially:

Jayne, Amy, Allen

Edie and Sandy

and

Rebecca

Acknowledgements

During my time at UCSF, I have benefited enormously from the knowledge and advice of many mentors. First, thanks to my advisor Andrej Sali. Andrej has taught me the importance of clearly defining a research question before beginning the long work of answering it. He has also helped me understand that, as method developers, we should focus on doing things right, rather than worrying too much about being first.

I have learned so much from all the members of the Sali lab, but I want to especially thank my two primary mentors, Keren Lasker and Riccardo Pellarin. Keren guided me through my rotation project and helped me understand structure modeling and the particular challenges of electron microscopy data. She also taught me some very cool algorithms. Riccardo taught me everything I know about Bayesian inference and helped me understand the connections between physics, probability theory, and data modeling. He also became a very good friend.

I could not have done my work without the computational infrastructure of the Sali lab (everything from the smooth running of our cluster to our lab software, IMP), so I want to thank Ben Webb and Daniel Russel. They (along with Keren, Riccardo, and a long list of other lab members) built the tools I use every day, taught me how to use them, answered my endless emails, and made me a better coder and scientist.

The Sali lab is full of brilliant people who each bring a unique perspective to the shared domain we all work in. They're also very silly. Thanks for all the good conversations over fine beverages to, in random order, Ilan, Adrian, Daniel 2, Avner, Dave, Ignacia, Sara, Kate, Javi, SJ, Jeremy, Pat, Ursula, Max, Dina, Elina, GQ, Seth, Shruthi, and Barak. Thanks also to Hilary, our lab manager

for many years, who helped me countless times and always had a good story to share.

Thanks to my committee members David Agard and Yifan Cheng. I could always count on them to bring new perspectives to the problems I worked on, ask challenging questions, and remind me that I am not developing methods for their own sake but for the advancement of the EM community. I also want to thank professors Matt Jacobson, who helped clarify my thesis proposal and gave me science and career advice many times; Ajay Jain, an early mentor of mine in his algorithms class; Joe DeRisi, who led the Team Challenges which got me so excited for research at UCSF; Tanja Kortemme, who among other things helped me give better talks; and Patsy Babbitt, who helped with my thesis proposal. I also want to thank my rotation mentor Eugene Oh, who was incredibly patient and helpful despite my lack of lab experience.

I want to thank my collaborators, especially Justin Kollman (formerly of the Agard lab), and Alex Zelter of Trisha Davis' lab, who were extremely helpful when writing the manuscript that forms Chapter 3 of this thesis. I also would like to thank Geir-Arne Fuglstad, who has been incredibly generous with his time helping me to understand spatial statistics.

Thanks as well to my funding sources, the NSF Graduate Student Research Fellowship, NIH training grant and R01, and the UCSF Discovery Fellowship.

It has truly been an honor to be a student at UCSF, a place where creativity and collaboration are valued and where virtually anyone is ready to chat with or help out a fellow scientist. This culture is maintained in part by the professors, but also by the program administrators. I want to especially thank Rebecca Brown, the biophysics student coordinator, who keeps the program thriving and who has been mind-bogglingly helpful to me many times over the years.

I am enormously grateful to my very good friends and classmates Erik Jonsson, Aram Avila-Herrera, Peter Cimermančič, and Jonathan Gable. Despite working on very different topics they were always ready to listen to the scientific problems that were driving me crazy, and usually offered creative solutions (or at least were ready with a miserable anecdote of their own to cheer me up!). Thanks also to my other classmates, Si-Han, Melanie, Zhiyuan, Laurens, and Zac. Ever since Bootcamp, our first classes, and Team Challenge, I have been enormously impressed by your drive and intelligence, and grateful for all the things you have taught me. Also, those early potlucks were really fun.

I also want to thank students of previous classes, especially David Booth, my earliest mentor at UCSF, from whom I inherited Bootcamp; thanks also to Genevieve Erwin and Henry Lin. Thanks as well to Kyle Barlow (Bootcamp is now in good hands). And a big thank-you to Jess, Max, Joe, Janet, Mélodie, Akhil, Kat, Serge, Bryant, Daniel, and Laura.

I am enormously grateful to my family. Their love, support, and patience through the years made graduate school much more bearable. Thanks to my mom, Jayne, who taught me to think creatively, and who always managed to be impressed by my updates no matter how dour they seemed to me. Thanks to my sister, Amy, who is always ready to spring a far-too-clever joke or turkey impression; she has also supported me through many tough times and, of course, navigating the aftermath of whatever stupid thing I just did. Thanks to my Uncle Steve, who has helped me in countless ways, including being a scientific mentor, and also being a good representative of my dad (Allen, who I would also like to thank) and reminding me that he would have been very pleasantly surprised by Amy and I doing so well. Thanks to Grandma Edie, who was always beaming with pride and who supported my education throughout the years, even tutoring me as a kid; her memory inspires me to be a better person. Thanks also to Grandma Sandy, who worked very hard to support my mother and her sisters, including taking care of me and Amy as kids, and whose memory inspires me to work hard.

Thanks also to my wife's family: Mary, Ric, Dottie, Emily, Katie, Paul, and Laura, who have all been incredibly supportive and welcoming to me.

Thanks finally to my wife Rebecca, whose help, support, and faith in me has made this thesis possible. I admire her for her brilliance, work ethic, commitment to good science, perceptiveness, humility, kindness, and incredible gratitude towards others. Her presence in my life has made every experience so much better. Even this.

Contributions

Chapter 3 of this thesis is a reprint of the material as it appears in:

C. H. Greenberg, J. Kollman, A. Zelter, R. Johnson, M. MacCoss, T. Davis, D. Agard, A. Sali. "Structure of γ-tubulin small complex based on a cryo-EM map, chemical cross-links, and a remotely related structure." *Journal of structural biology* 194, no. 3 (2016): 303-310.

The co-authors J. Kollman, A. Zelter, R. Johnson, M. MacCoss, T. Davis, and D. Agard are collaborators who provided the experimental data (including electron microscopy maps and chemical cross-links) which were used for modeling. Charles performed all the modeling tasks himself and wrote the manuscript with input and supervision from T. Davis, D. Agard, and A. Sali.

Appendix 1 of this thesis is a partial reprint of the material as it appeared in:

P. Robinson, M. Trnka, R. Pellarin, C. Greenberg, D. Bushnell, R. Davis, A. Burlingame, A. Sali, R. Kornberg. "Molecular architecture of the yeast Mediator complex." *Elife* 4 (2015): e08719.

The co-authors P. Robinson, M. Trnka, and R. Pellarin directed and supervised the research for this publication. Charles helped develop the EM fitting approach used in the publication, provided software support, and contributed to the manuscript.

Appendix 2 of this thesis is a reprint of the material as it appeared in:

J. Kollman, C. Greenberg, S. Li, M. Moritz, A. Zelter, K. Fong, J. Fernandez, A. Sali, J. Kilmartin, T. Davis, D. Agard. "Ring closure activates yeast γTuRC for species-specific microtubule nucleation." *Nature structural & molecular biology* 22, no. 2 (2015): 132-137.

The co-author J. Kollman directed and supervised the research for this publication. Charles performed modeling tasks and contributed to the manuscript.

Abstract

Advances in electron microscopy (EM) allow for structure determination of large macromolecular machines at increasingly high resolutions. A key step in this process is interpreting the EM density map with structural models of maximal accuracy and optimal precision. Model precision should be determined by the uncertainty in the experimental data; however, current methods only set uncertainty in an *ad hoc* manner with arbitrary weight terms. Thus, there is a need for more objective methods.

In Chapter 2, I present a novel Bayesian approach to modeling macromolecular structures based on EM density maps. The key advancement is the development of a scoring function that uses the local uncertainty of the density map to set the data weight and allows for correlation between neighboring density values. Unlike traditional approaches, the score does not require an expert user to set arbitrary parameters. I assessed the accuracy of models generated by this approach with a set of experimentally-derived, previously-published EM data of macromolecular complexes at varying resolutions from 3 to 6Å. I found that this approach leads to higher fluctuations in the model ensemble in locations with higher local uncertainty, and obtained accurate ensembles for a 3.2Å resolution map of Trpvl and 3.4Å and 5.4Å resolution maps of γ-secretase.

In Chapter 3, I present models of the γ-tubulin small complex in two functional states based on a challenging data set consisting of low-resolution EM density maps and a remotely related structure. Here, I used traditional scoring techniques, but extensively sampled alignments and conformations in order to ensure that the model ensemble reflected the uncertainty in the data. The resulting models form a tight cluster for each state and were consistent with a set of newly reported chemical cross-links. Comparing the two states, I found significant structural differences and predict stabilizing interactions of the two states. The work in this chapter shows the difficulties of traditional modeling and serves as motivation for the methods developed in Chapter 2.

Both approaches are incorporated into the open source *Integrative Modeling Platform* (IMP) package, enabling integration with multiple other data types and usage of myriad sampling and analysis tools.

Contents

Chapter 1: Introduction 1

Uncertainty in EM data 2

Error mitigation during reconstruction 3

Quality metrics 5

Approaches for building models with EM data 6

Outlook for new approaches 9

References 10

Chapter 2: Bayesian modeling of macromolecular structures based on electron microscopy density maps **14**

Introduction 15

Approach 16

Overview 16

Bayesian scoring function 16

Gaussian process model 17

Approximating the Gaussian process with a Gaussian Markov Random Field . . . 21

Calculation notes and caveats 26

Sampling procedure 27

Results 27

Local resolution drives local fluctuations 28

Tolerance of model-data violations is controlled by estimated uncertainty 28

Sampling the model posterior leads to more accurate ensembles 30
Discussion 35
Addendum 1: Non-stationary precision matrix 37
Addendum 2: Derivatives 41
References 42
Chapter 3: Structure of γ-tubulin small complex based on a cryo-EM map, chemical cross-links, and a remotely related structure 45
Introduction 47
Approach 48
Closed-state model building 48
Open-state model building 50
Model evaluation 51
Cross-linking of recombinant γTuSC and mass spectrometry analysis 51
Results 52
Initial model accuracy and coverage 52
Estimating model precision and accuracy 53
Comparing the open and closed states 57
Discussion 62
References 64
Appendix 1: EM Methods for the molecular architecture of the yeast Mediator complex 67
Appendix 2: Ring closure activates yeast γTuRC for species-specific microtubule nucleation 73

List of Figures

Chapter 2: Bayesian modeling of macromolecular structures based on electron microscopy density maps

1 Samples from the likelihood 25

2 Effect of local resolution on local fluctuations. 29

3 Correlation range sets the data tolerance 31

4 Score is more tolerant for lower quality data 32

5 Sampling structures of Trpv1 34

Chapter 3: Structure of γ-tubulin small complex based on a cryo-EM map, chemical cross-links, and a remotely related structure

1 Method overview 49

2 Data overview 54

3 Progress of modeling of each state 55

4 Modeling results 56

5 Cross-linking data analysis 58

6 Model comparison 60

List of Tables

Chapter 3: Structure of γ-tubulin small complex based on a cryo-EM map, chemical cross-links, and a remotely related structure

1 Contact overlap 61

Chapter 1

Introduction

The aim of structural biology is to create models of the structure and dynamics of macromolecular machines. Modeling these large, complex systems will help us understand how they function, how they evolved, and perhaps how they can be modified or even designed. The best way to determine these models is to utilize all available information simultaneously, including experimental data, physical principles, statistical inference, and intuition, in a modeling approach known as Integrative Structural Biology (Sali et al. 2015).

The process of Integrative Structural Biology has 4 stages: gathering information, designing model representation and evaluation, sampling good-scoring models, and analyzing the ensemble of models produced by sampling (Alber et al. 2008; Fernandez-Martinez et al. 2012; Lasker et al. 2010; Lasker et al. 2012; Schneidman-Duhovny et al. 2014). For fixed data input, two key factors shape the success of this process: choosing an accurate scoring function to encode agreement to the gathered information, and thoroughly sampling the space of models that score well against the data. Demonstrating the accuracy of the score and the completeness of the sampling are particularly needed for communicating the results of a structural study to the scientific community, as these markers indicate that the model-generating process was as accurate and precise as possible, given the input data.

The goal of this thesis is to develop procedures to generate accurate ensembles of structures given data from cryo-electron microscopy (cryo-EM) density maps. Below, we discuss the challenges of incorporating experimental data during modeling, investigate the specific sources of uncertainty in EM density maps, and discuss other approaches to generating good models from these data. In Chapter 2 I present a novel, probabilistic scoring approach that explicitly models the uncertainty in cryo-EM data and uses it to set the tolerance for data violations. In Chapter 3 I focus on ensemble generation for a challenging application with significant biological implications.

Uncertainty in EM data

Building models of a biological system consistent with all input data is a significant challenge in biology. Four data issues can lead to uncertainty about the model: (i) sparseness (lack of total coverage by the data); (ii) ambiguity (observations in the data do not map 1:1 to system components); (iii) incoherence (data points are not assignable to a specific state of the system); and (iv) random and systematic errors that affect the data values directly (Schneidman-Duhovny et al. 2014). During modeling, we should identify how these types of uncertainty arise from the experiments and create ensembles of models that indicate, to the degree possible, how much can be said about the system due to those uncertainties.

In the case of electron microscopy density maps, some of the above sources of uncertainties are present: (i) sparse sampling of projection angles by the 2D image set (which may even prevent reconstruction); (ii) ambiguity is only an implicit problem for EM density maps, as the data is applicable instead to the system as a whole; (iii) EM images may represent a mixture of multiple states; and (iv) random error results from noise in the images and systematic error results from mistakes in the correction of the contrast transfer function (CTF, defined below) (Penczek 2010) as well as the alignment of the 2D images (Scheres 2012b; Grigorieff 1998).

Error mitigation during reconstruction

2D image model and corrections

The first step towards reconstruction is to correct for distortions created by the microscope. The basic image formation model in electron microscopy is:

$$g_n(\mathbf{x}) = \mathrm{psf}_n(\mathbf{x}) * e_n(\mathbf{x}) * \left[\mathbf{R}_n\mathbf{r} + m_n^S(\mathbf{x})\right] + m_n^B(\mathbf{x}) \tag{1.1}$$

where g_n is a 2D image; psf_n is the point spread function of the image (the inverse Fourier transform of the CTF, which is affected by the voltage of the microscope and its defocus as well as other parameters); e_n is an envelope function in real space; $\mathbf{R}_n$ is a Radon transform computing line integrals of the molecule in the direction of the image plane; and m^B and m^S are background and scattering noise, respectively (Penczek 2010).

The model (1.1) is normally written in reciprocal space, where the convolutions become multiplications, and the 2D frequency components of the image inform a central section of the 3D frequency components of the 3D density (Penczek 2010). Usually the CTF can be determined from the settings of the microscope, but a more exact procedure fits parameters of the CTF directly to the micrograph (Mindell et al. 2003; Rohou et al. 2015). The envelope function, which truncates high frequency information due to various deficiencies in the experiment including radiation damage, is usually approximated as a Gaussian function and can in some cases be fitted to the data (Scheres 2016). With these parameters known, the images can be corrected to reduce distortion (although the CTF does reduce the signal to zero at some frequencies, this can be corrected by taking the data at multiple defocus settings, (see Penczek 2010)).

Recently, the quality of data sets was significantly improved by developments in detector technology and image processing (see comprehensive reviews Cheng et al. 2015; Bai et al. 2015a). Direct electron detectors have replaced film and CCDs due to their much improved quantum efficiency (a measure of how well a medium converts incident radiation to photons) (Li et al. 2013b; Ruskin et al. 2013). Image processing improvements are twofold: first, beam-induced motion was reduced by capturing EM "movies" (multiple short-exposure snapshots) instead of longer-exposure images (Brilot et al. 2012; Campbell et al. 2012; Li et al. 2013a). The multiple snapshots can dramatically improve the signal:noise ratio when aligned, leading to reconstructions based on far fewer images than before (*e.g.,* reconstruction of the ribosome at 4.5Å using only 30,000 images was achieved in Li et al. 2013a). The second image processing improvement was in the treatment of heterogeneity and uncertainty during 3D reconstruction, discussed below.

Creating EM density maps from 2D images

Successful 3D reconstruction can be hindered by heterogeneity in the sample. The best way to reduce compositional heterogeneity is with improved purification; conformational heterogeneity may be reduced with biochemical techniques (Cheng et al. 2015). Furthermore, if the states are sufficiently distinct, the images may be amenable to clustering techniques—these include multi-reference alignment, outlined in Frank 2006; likelihood-based classification (Sigworth 1998; Scheres et al. 2007); and bootstrapping EM images to identify substates within the data (Penczek et al. 2011). If only a part of the molecule is heterogeneous, one can even focus the classification in that region, leaving the most alignable regions fixed (a recent example is Bai et al. 2015b). If these techniques fail to separate the images into states, then likelihood-based tools such as RELION (Scheres 2012a) and FREALIGN (Grigorieff 1998) can sometimes still succeed by performing classification simultaneously with reconstruction.

Both RELION and FREALIGN use the same general approach: assume the 3D frequency components of each state (number of states specified at the beginning) are independent random variables, and find their optimal values based on the data and prior knowledge. Each tool uses the same core likelihood function to evaluate how well the 3D model matches the 2D frequencies of each image (also assumed independent). And, finally, both tools use expectation-maximization to optimize the 3D frequencies of each state as well as the noise in each image.

The crucial difference is in how the two methods treat "nuisance parameters" including the orientation of each image and what state it belongs to. Here, a nuisance parameter is a quantity whose value is not of direct interest, but is needed for the likelihood calculation. FREALIGN finds optimal values for these parameters, reasoning that the likelihood is sharply peaked around orientations and assignments that are accurate—thus FREALIGN is a "maximum likelihood" (ML) method.

In contrast, RELION makes no such assumption, instead marginalizing over values of these parameters and thus propagating their uncertainty to that of the final map(s). This procedure, called "maximum *a posteriori*" (MAP), is more objective but can be slower than FREALIGN. However, a key benefit to considering uncertainty in angle and state assignment is that RELION can benefit from setting priors on the maps' 3D frequency components to keep them close to zero—the Bayesian equivalent of regularization. This treatment imposes smoothness on the density maps, so that high-frequency components only move from zero if the uncertainty in the data is minimized. Thus, RELION has a more rigorous uncertainty-reduction approach. In Chapter 2, I utilize the expected smoothness of the EM density map to write the expected covariance in real space; this formulation enables evaluation of the significance of an atomic model built into the map.

Quality metrics

Evaluating the effect of these uncertainties on the quality of the EM density map is difficult, as the errors principally affect the underlying images, but the density map is a model generated from those images. A common metric for quality of a 3D reconstruction is "resolution," defined by the "gold standard" Fourier shell correlation procedure (Scheres et al. 2012). In this process, the EM images are divided into half-sets and reconstructed independently. The two "half-maps" are then compared in reciprocal space by the correlation at each frequency shell:

$$FSC = \frac{F_1 F_2^*}{\sqrt{\sum |F_1|^2 \sum |F_2|^2}} \tag{1.2}$$

where FSC is the correlation at one resolution shell, F_1 and F_2 are the Fourier transforms of the half maps of the two densities, and the sums run over all the values at each shell.

Starting at low frequencies, the two maps are expected to agree; at higher frequencies, the individual noise of each half-map become more significant. The resolution is usually reported as the frequency at which the correlation drops below a specified threshold. A common choice is 0.143, which was the result of calculating the expected FSC between half-sets when the full data set contains just enough information to support allowing one additional frequency shell to be considered signal (Rosenthal et al. 2003).

Recognizing that the uncertainties may vary throughout the density map, a method for evaluating "local resolution" was recently published (Kucukelbir et al. 2014). In this approach, 3D sinusoids of different size are fitted at each voxel of the density map. The largest sinusoid that correlates well with the density (according to a probabilistic measure) gives the local resolution at that point. While this approach is useful for evaluating the data quality in different regions, one must be careful about using the local resolution values themselves as estimates of *local uncertainty*. Feature sizes in real space are defined by a combination of the inherent shape of structural elements as well as the degree of uncertainty. In contrast, the gold standard FSC is more rigorously defined to explicitly calculate the overall degree of uncertainty.

Approaches for building models with EM data

Many techniques have been developed to build models into these density maps, starting with structures of related proteins or the same proteins in different states. These "flexible fitting" methods are local optimizers relying on a weighted scoring function, generally written as:

$$E_{total}(X) = w_{data}E_{data}(X) + E_{phys}(X) \tag{1.3}$$

where X is the coordinates of the model and w_{data} controls the importance of the data with respect to the force field, which often needs to be set by an expert user. I discuss issues with this formulation of scoring functions later in this section.

A commonly used score for evaluating fit quality (E_{data} above) is the cross-correlation coefficient, a measure of shape overlap between the target density and a simulated map:

$$CCC = \frac{\sum_{i,j} \rho_i^{EM} \rho_j^P}{\sum_i (\rho_i^{EM})^2 \sum_j (\rho_j^P)^2} \tag{1.4}$$

where ρ_i^{EM} is the EM density value at voxel i, ρ_j^P is the density of the theoretical map (computed by convolving the atoms of the model with a Gaussian function with variance calculated from the resolution) at voxel j, and the sums run over all the voxels. This score is the target function of Flex-EM (Topf et al. 2008). Other techniques include Molecular Dynamics Flexible Fitting (MDFF), which avoids map simulation by converting the target density map into a vector field based on the density gradient (Trabuco et al. 2008); and Direx, which uses deformable elastic networks (Wang et al. 2012).

Each of these tools can generate useful structural models; however, as local optimizers, none are specifically designed with the goal of sampling the ensemble of structures that are consistent with an EM density map. In Chapter 3, I used MDFF to solve a challenging modeling problem at low resolution. As I describe, MDFF did not automatically find a wide range of solutions; I had to repeatedly randomize the starting configuration (which was generated from a related molecule) in order to obtain enough structures to claim that the precision of the model ensemble reflected the uncertainty in the data.

A further limitation of the above approaches is that they each require *ad hoc* choices for weighting the experimental data against prior knowledge (w_{data} in Equation 1.3). These techniques do not explicitly consider the information content of the data (*i.e.*, the number of images that were used during reconstruction or the resolution of the map) nor the variability in that information (*i.e.*, the presence of higher or lower resolution regions of the data (Kucukelbir et al. 2014)). Rather, the user must make their own assessment when choosing the relative weight, and this can dramatically affect resulting models. For example, in Direx the user sets a parameter which controls the extent to which the mean positions of an elastic network of the target structure is allowed to shift during sampling; good results were only obtained for specific values of this parameter, and it is not clear whether it can be generalized between systems (Wang et al. 2012). This deficiency limits the accuracy of the produced model and the accuracy of its precision estimate. Thus, there is a need for a method that uses the data uncertainty to automatically weight the EM density map. Such weighting is the motivation for the method developed in Chapter 2.

For very high resolution systems (<4Å), it has become common to use tools developed for interpreting X-ray crystal structures, such as Refmac (Murshudov et al. 2011). In this approach, as in RELION and FREALIGN, the 3D frequency components of the density map are considered independent random variables with Gaussian noise. The score calculates the likelihood of the residual values between the observed frequencies and theoretical ones based on the current model. The noise terms can be set by comparing frequencies of half-sets of the images (Murshudov 2016). A limitation of this approach, however, is in the assumption of independence of the frequency components, which ignores the statistical effects of locally-varying uncertainty. Furthermore, as a maximum *a posteriori* approach, Refmac does not create ensembles of solutions.

Lastly, two new techniques incorporate local uncertainty into the model-building process. Rosetta incorporated a score that is a weighted combination of global and local cross-correlation coefficients and other score terms (where local CCC is just Equation 1.4 evaluated in the vicinity of a residue) (Wang et al. 2016). The weights were set by calibration on a different system from the target density map; thus, it is unclear how generalizable the values are. Additionally, a variant of MDFF was recently developed to improve convergence during sampling by (i) using "resolution exchange" to simultaneously run MDFF at a variety of map weights and (ii) scaling the map weight in different regions based on the local resolution (Singharoy et al. 2016). The results indicate higher fluctuations in lower-resolution regions; however, the new techniques are aimed at encouraging more sampling to escape local minima, rather than trying to score more accurately based on the data quality.

Outlook for new approaches

The explosion in availability of high-resolution EM density maps has been termed "The Resolution Revolution" (Kühlbrandt 2014). Improvements in microscopy and detector technology, image processing, and computational power have combined to make it possible to create models of biological macromolecules that are more accurate and precise than ever before. A small collection of these results includes Trpv1 at 3.1Å (Liao et al. 2013) and more recently at 2.9Å (Gao et al. 2016); β-galactosidase at 3.3Å (Bartesaghi et al. 2014) and 2.2Å (Bartesaghi et al. 2015); γ-secretase at 3.4 Å (Bai et al. 2015c); 20S proteasome at 3.3Å (Li et al. 2013a) and 2.8Å (Campbell et al. 2015); ryanodine receptor at 4.9Å (Wei et al. 2016); and the structures of a eukaryotic motor protein in multiple states from 6.9-8.3Å (Zhao et al. 2015).

At these near-atomic resolutions, it becomes possible to explore many interesting questions about the function and dynamics of macromolecules, particularly since cryo-EM data is obtained essentially for macromolecules in solution. To explore the implications of these data sets, it is more important than ever to create models with precision that reflects the uncertainties in the data. For new methods, this will require rigorously evaluating the uncertainty created by the images and using that information to set tolerances between the model and the data.

References

Alber, F. et al. 2008. “Integrating diverse data for structure determination of macromolecular assemblies.” *Annu Rev Biochem* 77:443–77.

Bai, X.-C., McMullan, G., and Scheres, S. H. W. 2015a. “How cryo-EM is revolutionizing structural biology.” *Trends in Biochemical Sciences* 40 (1): 49–57.

Bai, X.-C. et al. 2015b. “Sampling the conformational space of the catalytic subunit of human γ-secretase.” *Elife* 4:e11182.

Bai, X.-c. et al. 2015c. “An atomic structure of human [ggr]-secretase.” *Nature*.

Bartesaghi, A. et al. 2015. “2.2 Å resolution cryo-EM structure of β-galactosidase in complex with a cell-permeant inhibitor.” *Science* 348 (6239): 1147–1151.

Bartesaghi, A. et al. 2014. “Structure of β-galactosidase at 3.2-Å resolution obtained by cryo-electron microscopy.” *Proceedings of the National Academy of Sciences* 111 (32): 11709–11714.

Brilot, A. F. et al. 2012. “Beam-induced motion of vitrified specimen on holey carbon film.” *Journal of structural biology* 177 (3): 630–637.

Campbell, M. G. et al. 2012. “Movies of ice-embedded particles enhance resolution in electron cryo-microscopy.” *Structure* 20 (11): 1823–1828.

Campbell, M. G. et al. 2015. “2.8 Å resolution reconstruction of the Thermoplasma acidophilum 20S proteasome using cryo-electron microscopy.” *Elife* 4:e06380.

Cheng, Y. et al. 2015. “A primer to single-particle cryo-electron microscopy.” *Cell* 161 (3): 438–449.

Fernandez-Martinez, J. et al. 2012. “Structure–function mapping of a heptameric module in the nuclear pore complex.” *The Journal of cell biology* 196 (4): 419–434.

Frank, J. 2006. *Three-dimensional electron microscopy of macromolecular assemblies: visualization of biological molecules in their native state*. Oxford University Press.

Gao, Y. et al. 2016. “TRPV1 structures in nanodiscs reveal mechanisms of ligand and lipid action.” *Nature*.

Grigorieff, N. 1998. “Three-dimensional structure of bovine {NADH}:ubiquinone oxidoreductase (complex {I}) at 22 Å in ice.” *J Mol Biol* 277 (5): 1033–46.

Kucukelbir, A., Sigworth, F. J., and Tagare, H. D. 2014. “Quantifying the local resolution of cryo-EM density maps.” *Nat Methods* 11 (1): 63–5.

Kühlbrandt, W. 2014. “The resolution revolution.” *Science* 343 (6178): 1443–1444.

Lasker, K. et al. 2010. “Integrative structure modeling of macromolecular assemblies from proteomics data.” *Molecular & Cellular Proteomics* 9 (8): 1689–1702.

Lasker, K. et al. 2012. “Molecular architecture of the 26S proteasome holocomplex determined by an integrative approach.” *Proceedings of the National Academy of Sciences* 109 (5): 1380–1387.

Li, X. et al. 2013a. “Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM.” *Nature methods* 10 (6): 584–590.

Li, X. et al. 2013b. “Influence of electron dose rate on electron counting images recorded with the K2 camera.” *Journal of structural biology* 184 (2): 251–260.

Liao, M. et al. 2013. “Structure of the TRPV1 ion channel determined by electron cryo-microscopy.” *Nature* 504 (7478): 107–112.

Mindell, J. A. and Grigorieff, N. 2003. “Accurate determination of local defocus and specimen tilt in electron microscopy.” *Journal of structural biology* 142 (3): 334–347.

Murshudov, G. N. et al. 2011. “REFMAC5 for the refinement of macromolecular crystal structures.” *Acta Crystallographica Section D: Biological Crystallography* 67 (4): 355–367.

Murshudov, G. 2016. “Refinement of Atomic Structures Against cryo-EM Maps.” *Methods in Enzymology*.

Penczek, P. A. 2010. “Image restoration in cryo-electron microscopy.” *Methods Enzymol* 482:35–72.

Penczek, P. A., Kimmel, M., and Spahn, C. M. 2011. “Identifying conformational states of macromolecules by eigen-analysis of resampled cryo-EM images.” *Structure* 19 (11): 1582–1590.

Rohou, A. and Grigorieff, N. 2015. “CTFFIND4: fast and accurate defocus estimation from electron micrographs.” *Journal of Structural Biology* 192 (2): 216–221.

Rosenthal, P. B. and Henderson, R. 2003. “Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy.” *Journal of molecular biology* 333 (4): 721–745.

Ruskin, R. S., Yu, Z., and Grigorieff, N. 2013. “Quantitative characterization of electron detectors for transmission electron microscopy.” *Journal of structural biology* 184 (3): 385–393.

Sali, A. et al. 2015. “Outcome of the First wwPDB Hybrid/Integrative Methods Task Force Workshop.” *Structure* 23 (7): 1156–67.

Scheres, S. H. W. 2012a. “A Bayesian view on cryo-EM structure determination.” *J Mol Biol* 415 (2): 406–18.

— . 2016. “Processing of Structurally Heterogeneous Cryo-EM Data in RELION.” *Methods in Enzymology*.

— . 2012b. “RELION: implementation of a Bayesian approach to cryo-EM structure determination.” *J Struct Biol* 180 (3): 519–30.

Scheres, S. H. W. and Chen, S. 2012. “Prevention of overfitting in cryo-EM structure determination.” *Nat Methods* 9 (9): 853–4.

Scheres, S. H. et al. 2007. “Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization.” *Nature methods* 4 (1): 27–29.

Schneidman-Duhovny, D., Pellarin, R., and Sali, A. 2014. “Uncertainty in integrative structural modeling.” *Current opinion in structural biology* 28:96–104.

Sigworth, F. 1998. “A maximum-likelihood approach to single-particle image refinement.” *Journal of structural biology* 122 (3): 328–339.

Singharoy, A. et al. 2016. “Molecular dynamics-based refinement and validation for sub-5 Å cryo-electron microscopy maps.” *eLife* 5:e16105.

Topf, M. et al. 2008. “Protein structure fitting and refinement guided by cryo-EM density.” *Structure* 16 (2): 295–307.

Trabuco, L. G. et al. 2008. “Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics.” *Structure* 16 (5): 673–83.

Wang, R. Y.-R. et al. 2016. “Automated structure refinement of macromolecular assemblies from cryo-EM maps using Rosetta.” *bioRxiv*: 050286.

Wang, Z. and Schröder, G. F. 2012. “Real-space refinement with DireX: From global fitting to side-chain improvements.” *Biopolymers* 97 (9): 687–697.

Wei, R. et al. 2016. “Structural insights into Ca2+-activated long-range allosteric channel gating of RyR1.” *Cell Research* 26 (9): 977–994.

Zhao, J., Benlekbir, S., and Rubinstein, J. L. 2015. “Electron cryomicroscopy observation of rotational states in a eukaryotic V-ATPase.” *Nature* 521 (7551): 241–245.

Chapter 2

Bayesian modeling of macromolecular structures based on electron microscopy density maps

Introduction

Building structural models of a biological system that are consistent with the myriad data describing them is one of the key challenges in biology. Integrative structure modeling addresses this challenge by utilizing all data simultaneously. The end result is an ensemble of structural models, the variation of which should be indicative of the input data and prior knowledge (Ward et al. 2013). In other words, the goal of modeling is to find an ensemble of models whose computed data match the experimental data within the experimental uncertainty.

Any modeling requires a scoring function for ranking and a conformational sampling algorithm guided by the function. Probabilistic techniques such as Inferential Structure Determination have been shown to obtain accurate results while limiting overfitting (Rieping et al. 2005). In this approach, the score is written as the posterior probability, defined as the distribution over possible models conditional on all evidence obtained from an experiment as well as prior knowledge. Accordingly, "weight" terms assigned (often arbitrarily) to experimental data are reinterpreted as parameters expressing the uncertainty of experimental information (*e.g.*, if the data have independent Gaussian noise, then the "weight" of a data point is the variance of the noise). Such "nuisance parameters" control the relative importance of the data and prior information. The immediate benefit of using probabilities is that all terms in the scoring function are commensurate and obey the rules of statistics and information theory, in stark contrast with traditional scoring functions where there is no rigorous procedure on how to combine the individual energy terms or assess the precision of the models (Habeck et al. 2006); in other words, the Bayesian posterior density is more "objective" than the traditional scoring functions.

Here, we develop the posterior density for the model given an EM density map. We consider the EM density map as a joint probability distribution over the values at each voxel. For the first time, we explicitly model the variability of the uncertainty in the EM density map as well as the spatial correlation of this uncertainty across neighboring voxels. This approach is more objective than traditional scoring because data weight is automatically set by the data quality. We achieve this goal using the formalism of Gaussian processes, taking advantage of recent developments in spatial statistics that make computation of the likelihood possible. We illustrated this method by applying it to a synthetic and real benchmark. The method has been made available to other researchers as part of the open-source *Integrative Modeling Platform* (Russel et al. 2012).

Approach

Overview

We represent our structure as a set of spheres with coordinates, radius, and mass corresponding to the primary sequence. We develop a scoring function that first simulates a "noiseless" theoretical density map on the same grid as the experimental map, and then calculates the residual voxel values between the two maps. The key advancements with this scoring function are in the consideration of expected value for these residuals, namely that they are not independent of each other. Specifically:

1. The amount of uncertainty in voxel values correlates with neighbors.

2. The degree of that correlation varies with location in the density map.

These assumptions are justified below as being caused by the errors in the underlying images, which also affects the measured local resoluion in the EM density map.

Bayesian scoring function

The posterior probability of model M given data D and prior knowledge I is

$$p(M|D,I) \propto p(D|M,I)p(M|I) \tag{2.1}$$

The model M consists of a structure X and unknown parameters σ, such as noise in the data. The prior $p(M|I)$ is the probability density of model M given I. The prior can reflect prior knowledge, such as excluded volume, statistical potentials, a molecular mechanics force field, and other experimental measurements. The likelihood function $p(D|M,I)$ is the probability density of observing data D given M and I, and is often expressed as the product of distributions over each observation based on a noise model of the experiment:

$$p(D|M,I) = \prod_i N(f_i(X), D_i) \tag{2.2}$$

where $f(X)$ is a forward model that predicts the data that would have been observed for structure X in an experiment without uncertainty and $N(f(X),D)$ is a noise model that specifies the probability density of the the deviation between the predicted and observed data.

For EM density maps, this formulation of the likelihood would consider each voxel as an independent measurement and write the likelihood as a product of univariate distributions centered around the theoretical voxel value at each point, modulated by the estimated noise. However, the voxel density values are not direct measurements; the raw data is contained in the pixel intensities in the underlying images (even these pixel values are not independent measurements because they are obtained by superposing and averaging multiple image snapshots). The density maps are actually *models* that were fit to the images, usually with iterative back-projection (Scheres 2012). As a consequence, all 2D images inform all parts of the map; thus, there may be statistical dependence between the voxels. Therefore, our goal is to find an appropriate functional form of the *joint* distribution of the voxel uncertainties.

Gaussian process model

The joint distribution

EM density maps consist of 3D voxels whose values correspond to the Coulomb potential of the molecule at the voxel location. In the following, we consider any location in the EM map as a random variable; our goal is to compute, as the likelihood, the distribution of these variables given the current structural model of the system and an estimate of uncertainty in the voxel values. We justify our approach by referencing the procedure used for 3D reconstruction in RELION (Scheres

2012), though the results are general for any reconstruction that employs priors or regularization to reduce high-frequency noise.

In RELION, the posterior density of 3D frequency components of the density map is sampled (Scheres 2012). Each image is taken as an independent set of observations. The likelihood of one image is given as the product of Gaussians over each 2D frequency component, with variances (one per 2D frequency component per image) that are also sampled. Importantly, model complexity is minimized by imposing a prior on the 3D frequency components, which expresses the assumption that the density map is perfectly smooth in the absence of data. Consequently, higher-frequency terms grow only with the coherence and coverage of the images.

We approximate this procedure as a pseudo-experiment in which the signal of the molecule is observed *directly*, with added white noise, but the result has high frequencies truncated (*i.e.*, a linear filter is applied). In real space one may write:

$$y(\mathbf{s}) = k_s(\mathbf{s}) * (m(\mathbf{s}) + x(\mathbf{s})) \tag{2.3}$$

where $k_s(\mathbf{s})$ is a point spread function, or kernel, that is allowed to vary in space (as some regions may have higher frequencies activated); $m(\mathbf{s})$ is the Coulomb potential of the molecular model; $x(\mathbf{s})$ is a white noise process; and $*$ is convolution. The first term, $\mathbf{V} := k * m$, is a deterministic function of the current model calculated by spreading the potential of each atom (see below). The second term, $k * x$, is a smooth random process.

With little loss of generality, we model $y(\mathbf{s})$ with a Gaussian process, defined as a collection of random variables, any finite number of which have a joint Gaussian distribution. A Gaussian process is defined entirely by its mean function (here $\mathbf{V}$) and a covariance function (not yet defined) (Rasmussen et al. 2006). We choose this process because it is quite general; many types of noise behavior can be modeled by parameters within the covariance.

The likelihood of the Gaussian process model is given by:

$$
\begin{aligned}
p(\mathbf{y}) &= \mathscr{N}(\mathbf{V},\mathbf{K}) \\
&= (2\pi)^{-\frac{Nvox}{2}}|\mathbf{K}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{V})^T\mathbf{K}^{-1}(\mathbf{y}-\mathbf{V})\right)
\end{aligned}
\tag{2.4}
$$

where $\mathbf{K}$ is the covariance matrix. Note that a stationary Gaussian process is one in which the covariance is only a function of distance, not of location (Rasmussen et al. 2006). By allowing our kernel to vary in space, we are developing an inherently *non-stationary* process for the residuals between the data $\mathbf{y}$ and simulated density map $\mathbf{V}$. In the next section, we present a model for the covariance $\mathbf{K}$ that is consistent with the linear model (2.3).

Real-space covariance set by the local uncertainty

Smoothed noise processes like (2.3) have been studied previously and are known as *process convolutions* (Higdon et al. 1999). It has been shown that if one defines the kernel at each point in space, and assuming it is finite, then a valid covariance is given by convolving the kernel with itself (Paciorek et al. 2006). For example, if one models the kernel as a Gaussian function with spatially-varying covariance matrix $\Sigma(\mathbf{s})$, then the covariance function is:

$$
k(\mathbf{x}_i,\mathbf{x}_j) = \mathscr{N}(\mathbf{x}_i-\mathbf{x}_j|\Sigma_i+\Sigma_j) \tag{2.5}
$$

While this formulation is appealing for its simplicity, the covariance matrix of the process is, in general, dense. Considering that the number of voxels in an EM map is on the order of 10^6, the matrix $\mathbf{K}$ has 10^{12} entries and cannot be decomposed with usual methods. Techniques to improve the tractability of the likelihood computation, such as Kronecker techniques (Gilboa et al. 2015), only work in the case of a stationary process. Other methods require some arbitrary assumptions to enforce sparseness in $\mathbf{K}$, as in the case of covariance tapering (Kaufman et al. 2008) or block-composite likelihoods (Eidsvik et al. 2014).

Instead of attempting to construct an appropriate and tractable non-stationary covariance, we simply assume that the process is locally stationary and avoid explicitly writing the kernel. That is, given some voxel $\mathbf{x}_o$ and neighborhood $N(\mathbf{x}_o)$, then for all points $\mathbf{x}_i, \mathbf{x}_j \in N(\mathbf{x}_o)$, $cov(\mathbf{x}_i, \mathbf{x}_j)$ is only a function of the distance $|\mathbf{x}_i - \mathbf{x}_j|$. Within these neighborhoods, we model the covariance with a generalization of the Gaussian distribution called the Matérn family of covariances:

$$r_{\mathbf{s}}(\mathbf{u},\mathbf{v}) = \frac{\sigma^2}{2^{\nu-1}}(\kappa||\mathbf{u}-\mathbf{v}||)K_\nu(\kappa||\mathbf{u}-\mathbf{v}||), \quad \mathbf{u},\mathbf{v} \in N(\mathbf{s}) \subseteq \mathbb{R} \tag{2.6}$$

which defines the covariance between points within local region $N(\mathbf{s})$ and where K_ν is the modified Bessel function of the second kind and order $\nu > 0$. The parameter κ is related to the "range" of the process $\rho = \sqrt{(8\nu)}/\kappa$, defined as the distance at which the correlation is 0.1 for any ν. The parameter ν sets the "smoothness" of the covariance; integer values correspond to the degree of differentiability at distance zero. For $\nu = 0.5$ the covariance reduces to $\sigma^2 \exp -\frac{|\mathbf{u}-\mathbf{v}|}{\rho}$, and as $\nu \to \infty$ the covariance approaches $\sigma^2 \exp -\frac{|\mathbf{u}-\mathbf{v}|^2}{2\rho^2}$. In practice it is usually difficult to determine ν, but we have found for EM density maps that $\nu = 2.5$ is a good balance between smoothness and computational tractability.

The final missing piece to this model is deciding how to set the covariance range ρ for each neighborhood. Recall that in (2.3) we modeled the data processing in EM density maps as a truncation of higher frequencies, and this effect can be spatially varying. A tool called ResMap has been developed to estimate this effect in different regions of the EM density map (Kucukelbir et al. 2014). This tool estimates "local resolution" by fitting sinusoids in the vicinity of each voxel and reporting the largest frequency that matches the data. We assume that the values provided by ResMap are proportional to the true ranges, and set:

$$\rho_i = \xi \cdot \mathrm{res}_i \tag{2.7}$$

where ρ_i is the local range, res_i is the local resolution given by ResMap, and ξ is a global nuisance parameter. Choosing the covariance range this way can be seen as an application of the Weiner-Khinchin theorem, which states that the autocorrelation of a stationary process is the Fourier dual of its spectral density (Rasmussen et al. 2006).

Approximating the Gaussian process with a Gaussian Markov Random Field

We have now (i) established the probabilistic framework of the residual map as a Gaussian process, (ii) demonstrated the expectation that within local regions, the covariance should be Matérn with the range parameter related to the local resolution, and (iii) realized that the matrix $\mathbf{K}$ needs to be decomposable to be useful for fitting a molecular model to the EM density map. Here, we show how it is possible to obtain such a covariance by only specifying local conditional probabilities as entries in the highly sparse precision matrix $\mathbf{Q} = \mathbf{K}^{-1}$.

Sparse approximation for the stationary case

A recent landmark result in spatial statistics showed that one can approximate a Gaussian process exhibiting the Matérn covariance (2.6) (and thus estimate our likelihood) with very high accuracy and vastly improved calculation speed using a Gaussian Markov random field (GMRF) (Lindgren et al. 2011). A GMRF is a set of normally distributed random variables that have a Markov property with respect to an undirected graph (*i.e.*, missing edges in the graph correspond to zeros in the precision matrix (inverse covariance) of the distribution) (Rue et al. 2005).

In general it can be difficult to find appropriate values for the precision such that it is positive definite. However, it was shown some time ago that solutions to the following stochastic partial differential equation (SPDE) exhibited the Matérn covariance (Whittle 1954):

$$(\kappa^2 - \Delta)u(\mathbf{s}) = \mathscr{W}(\mathbf{s}), \qquad s \in \mathbb{R}^d, \kappa > 0 \tag{2.8}$$

where $\Delta = \nabla \cdot \nabla = \sum_d \frac{\partial^2}{\partial d^2}$ is the Laplacian; κ is the same as in the Matèrn covariance, related to the covariance range via $\rho = \sqrt{8\nu}/\kappa$; and $\mathscr{W}$ is a white noise process defined by $\int_V \mathscr{W} dV = |V|$. The degree of smoothness is set by the value of ρ.[1]

[1] To understand why solutions to the SPDE are locally smooth, consider segmenting the space into sub-regions. The SPDE requires that, on average, $\kappa^2 u$ in a region is equal to the sum of the second derivatives of u in that region.

The next step in (Lindgren et al. 2011) was to replace the continuous process $u(s)$ with a discrete process $\mathbf{u}$ on a grid or a triangulation, followed by solving (2.8) with the finite element method (*i.e.*, on each grid volume separately). A key approximation is that the neighborhoods of grid points define their conditional dependencies; that is, a voxel is independent of far-away voxels given average values of its neighbors. This is not a particularly severe assumption for a continuous process, as discussed in (Simpson et al. 2012).

As a consequence of this approximation, on the discrete space the operator $(\kappa^2 - \Delta)$ can be written as a matrix $\mathbf{B}$ and the SPDE reduces to:

$$\mathbf{B}(\kappa^2)\mathbf{u} \sim \mathscr{N}(0, \mathbf{I}) \tag{2.9}$$

where the white noise process $\mathscr{W}$ is now just a set of independent normally distributed random variables at each voxel (Fuglstad et al. 2013). The entries of the highly sparse matrix $\mathbf{B}$ are easy to calculate; for a row i, the only non-zero columns entries are the those corresponding to the neighbors of grid point i. The actual values are calculated in Addendum 1. Following directly from (2.9), the approximated likelihood of the spatial process $\mathbf{u}$ is given by:

$$\begin{aligned} \pi(\mathbf{u}) &= \exp\left[-\frac{1}{2}\left(\mathbf{B}(\kappa^2)\mathbf{u}\right)^T\left(\mathbf{B}(\kappa^2)\mathbf{u}\right)\right] \\ &= \exp\left[-\frac{1}{2}\mathbf{u}^T\mathbf{Q}(\boldsymbol{\kappa}^2)\mathbf{u}\right] \end{aligned} \tag{2.10}$$

where $\mathbf{Q}(\kappa^2) = \mathbf{B}(\kappa^2)^T\mathbf{B}(\kappa^2)$ is the precision matrix.

To summarize, the neighborhood graph defines the independencies of the voxels, and we have solved the discrete SPDE (2.9) by calculating the precision matrix; thus, we have approximated a stationary Gaussian process (*i.e.*, (2.4) with entires of $\mathbf{K}$ only a function of distance) using a tractable GMRF. We conclude this section by extending this approximation to a higher-differentiability SPDE and then the full non-stationary process.

Thus for small values of κ (large values of the range ρ), the second derivatives are very small, and u is more closely related to neighboring region values. In contrast, large values of κ (small ranges ρ) require large second derivatives and so u changes faster. Thus, κ controls the smoothness, and the SPDE is expressing a kind of conservation law.

Higher order SPDE

The SPDEs presented in (2.8) and (2.9) lead to a stationary 3D Matérn covariance with order $\nu = 0.5$. Recall that the parameter ν sets the differentiability of the covariance, with $\nu \to \infty$ corresponding to the infinitely differentiable Gaussian function. As mentioned above, we found that a higher differentiability of the covariance was required for good results with EM density maps (data not shown). These values can be achieved, even with fractional differentiability, by taking the operator of the SPDE (2.8) to power $\alpha/2$ where $\alpha = \nu + d/2$. This is necessary when the ranges of interest are low. Extending the solution to second order simply requires expanding the set of neighbors; the solution is $\mathbf{Q}(\kappa^2) = \mathbf{B}(\kappa^2)^T\mathbf{B}(\kappa^2)^T\mathbf{B}(\kappa^2)\mathbf{B}(\kappa^2)$

Non-stationary extension

The final missing piece is the fact that the SPDE (2.8) assumes uniform range, but the Gaussian process model (2.4) is non-stationary (though we assume it is locally stationary). Luckily, it was shown that the SPDE can be extended to a non-stationary process with very little additional computational burden (Lindgren et al. 2011). More recently, this result was extended to allow non-stationary and non-isotropic covariance structure (Fuglstad et al. 2013). Here we use a non-stationary but isotropic variant of the SPDE:

$$(\kappa^2(\mathbf{s}) - \Delta)^2(u(\mathbf{s})) = \mathscr{W}(\mathbf{s}) \tag{2.11}$$

where $\kappa(\mathbf{s})$ controls the locally-varying range. Assuming the range parameter does not change too quickly (equivalent to our assumption that the process is locally stationary), the covariance is still Matérn. We solve the SPDE for 3D cases in Addendum 1, calculating the precision matrix $\mathbf{Q}$ explicitly.

In conclusion, we have approximated the Gaussian process model with a locally Matérn covariance and spatially-varying range set by the local resolution via (2.7). The solution to the non-stationary SPDE (2.11) exhibits the desired local behavior while also being computationally tractable due to the explicitly sparse terms of the precision matrix. The full covariance matrix **K** never has to be computed or inverted.

As a test, we can draw random samples from the Gaussian process likelihood (2.4) with a random number generator and using the Cholesky decomposition of the covariance (or precision) matrix to smooth out the random numbers (Rasmussen et al. 2006; Rue et al. 2005). We drew random samples in 2D and 3D using the non-stationary precision matrix via solutions to the SPDE (2.11) and displayed them in **Figure 1**; these samples represent random *residuals* between a model and a density map. Note that in regions with longer range, the sample is "clumpier." This effect indicates that voxels in regions of high uncertainty (long range) are less informative than those in regions of low uncertainty (short range).

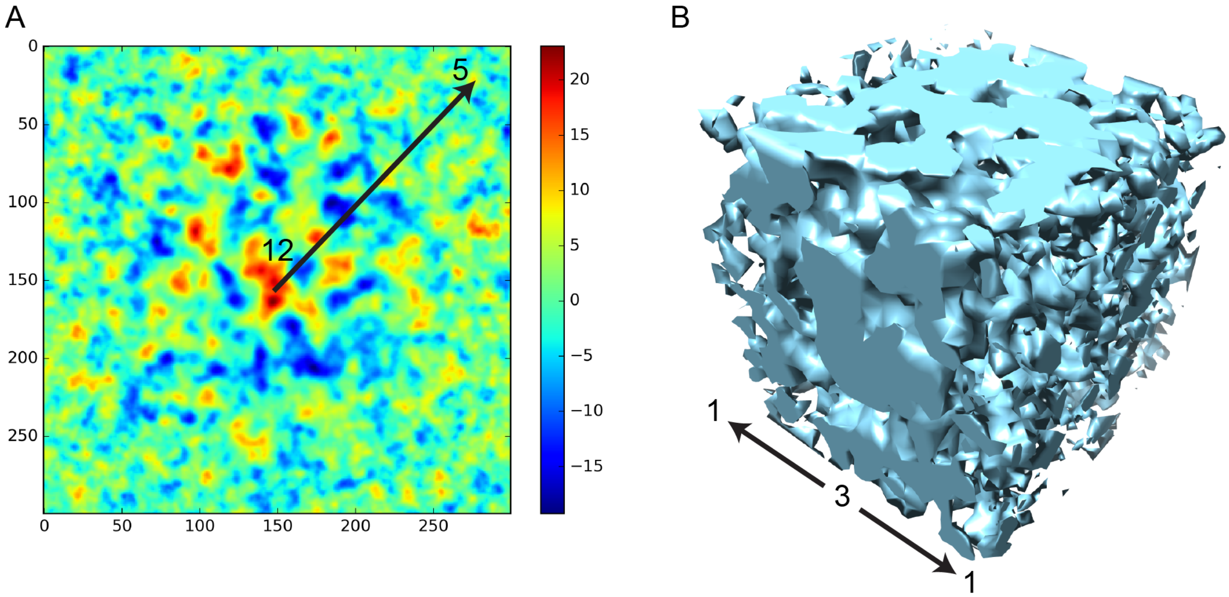


Figure 1: Samples from the likelihood. (A) Here, we created a non-stationary 2D distribution by setting the covariance to be large at the center and small at the edges (arrow indicates the smooth progression of range) with periodic boundary conditions. Color is the random number in each bin. (B) Same calculation but in 3D, with the range only changing along the indicated axis. All units are in pixels/voxels.

Calculation notes and caveats

Determinant

Calculating the likelihood (2.4) requires the computation of the determinant of $\mathbf{K}$. This computation usually requires the calculation of the Cholesky decomposition of $\mathbf{K}$ which is difficult for the typically large sample sizes. Therefore we used a method for estimating the log-determinant with a short Monte Carlo sampling procedure ($\sim$30 seconds for a 150x150x150 system) (Han et al. 2015).

Correcting the marginal variance

A bigger challenge is the fact that the non-stationary covariance model creates a non-uniform marginal variance that scales with the correlation distance. This scaling may be undesired. To remove this effect, one must correct $\mathbf{Q}$ by the diagonal of its inverse, the covariance:

$$\mathbf{Q}_{\text{fix}} = D_i \mathbf{Q} D_i \tag{2.12}$$

where D_i is the diagonal of $\mathbf{K} = \mathbf{Q}^{-1}$. The computation of the diagonal of the inverse (the covariance matrix) is a difficult problem; the straightforward approach is to solve $\mathbf{Qx} = \mathbf{I}$ where $\mathbf{I}$ is the identity matrix. However, this approach is intractable due to the large number of data points. We have attempted to employ a method designed for very large GMRFs (Malioutov et al.). The idea is to solve a reduced rank problem $\mathbf{Qx} = \mathbf{A}$ where $\mathbf{A}$ is an $N_{vox} \times M$ matrix and $M << N_{vox}$, that assigns each voxel to a "color." The number of colors, and their assignments, are chosen in such a way that voxels of the same color are far beyond the maximum correlation distance of the GMRF. The advantage is that $\mathbf{Qx} = \mathbf{A}$ is a much faster problem to solve using sparse iterative solvers. While this successfully corrected problems up to size $50 \times 50 \times 50$, we have not yet successfully scaled it to larger sizes. In the future we hope to parallelize this approach to handle larger systems.

The simulated density map

In the likelihood (2.4), the sampled density map is given as the deterministic function $\mathbf{V} = k * \mathbf{m}$, the convolution of the current model atom positions (as point particles with mass) with the spatially-varying kernel function. However, in the GMRF approach we avoid explicitly calculating the kernel function, instead directly constructing a spatial process with the desired local covariance. For complete accuracy, the sampled density map $\mathbf{V}$ should be obtained by solving a non-stochastic version of (2.11), replacing the noise term with a vector of mass values of the model in each voxel. This would be impractical since the simulated density map needs to be calculated every time the likelihood is evaluated. Instead, we simply convolve $\mathbf{m}$ with a spatially-varying Gaussian function with standard deviation set to 0.425 times the local resolution, as suggested in (Topf et al. 2008). In the future, we plan to explore more ways of simulating the density map that are consistent with the SPDE approach.

Sampling procedure

We can sample the model with a molecular dynamics (MD) algorithm in which the potential energy of the system is replaced by the model posterior, consisting of a physics-based force field prior and the data likelihood (2.4). The calculation of derivatives require little additional work (see Addendum 2). While currently we keep the resolution multiplier ξ fixed, in the future we will employ a Monte Carlo sampler (random number generator) and sample this value together with model coordinates in a Gibbs sampling scheme (Rieping et al. 2005).

Results

The goals of this approach are to maximize the accuracy and optimize the precision of the ensemble of structures sampled from the model posterior. Criteria for achieving these are: the model ensemble should be centered around the true structure, as measured by its agreement to a higher-resolution reference data set; and the size of the ensemble should closely match the uncertainty of the data, as measured by its local resolution.

To evaluate our progress towards these goals, we test that (i) sampling the model posterior leads to local fluctuations of the model that are driven by the uncertainty; (ii) the data likelihood (2.4) correctly sets the tolerance of disagreement between the model and the data based on the uncertainty (local resolution); and (iii) good structures are found during sampling when starting far away, resulting in the size of the sampled ensemble reflecting the uncertainty of the target density map.

Local resolution drives local fluctuations

We first explored how setting the local resolutions of a density map manually in a simulated setting would affect the size of fluctuations (measured by the Cα root mean square fluctuation, or RMSF) during sampling (**Figure 2**). As local resolution is used to define the covariance range (and thus the effective "weight" of the data likelihood in different regions) we expected that, no matter how the resolution values are distributed, higher fluctuations should correlate with lower resolutions. We simulated two different maps with very different local resolutions, and performed short MD simulations. The result is that local resolution can control RMSF in a predictable fashion, and that this effect can override any inherent variable flexibility of the structure.

Tolerance of model-data violations is controlled by estimated uncertainty

Next we explored how modifying the local resolution parameters affected the magnitude of data likelihood (2.4), and its ranking of good structures. We expected that lowering input resolution (*i.e.*, increasing the covariance range) should rank a set of fixed models more closely together, indicating that tolerance is being correctly set by the resolution.

Recall that the local covariance range is set as a constant multiple (within one data set) of the local resolution: the parameter ξ is the multiplier. In this test we varied the value of ξ to see its effect on tolerance (**Figure 3**). For this test we used a real data set, γ-secretase at 5.4Å resolution (EMDB: 2678, Lu et al. 2014). As ξ increases, the correlations become longer, causing

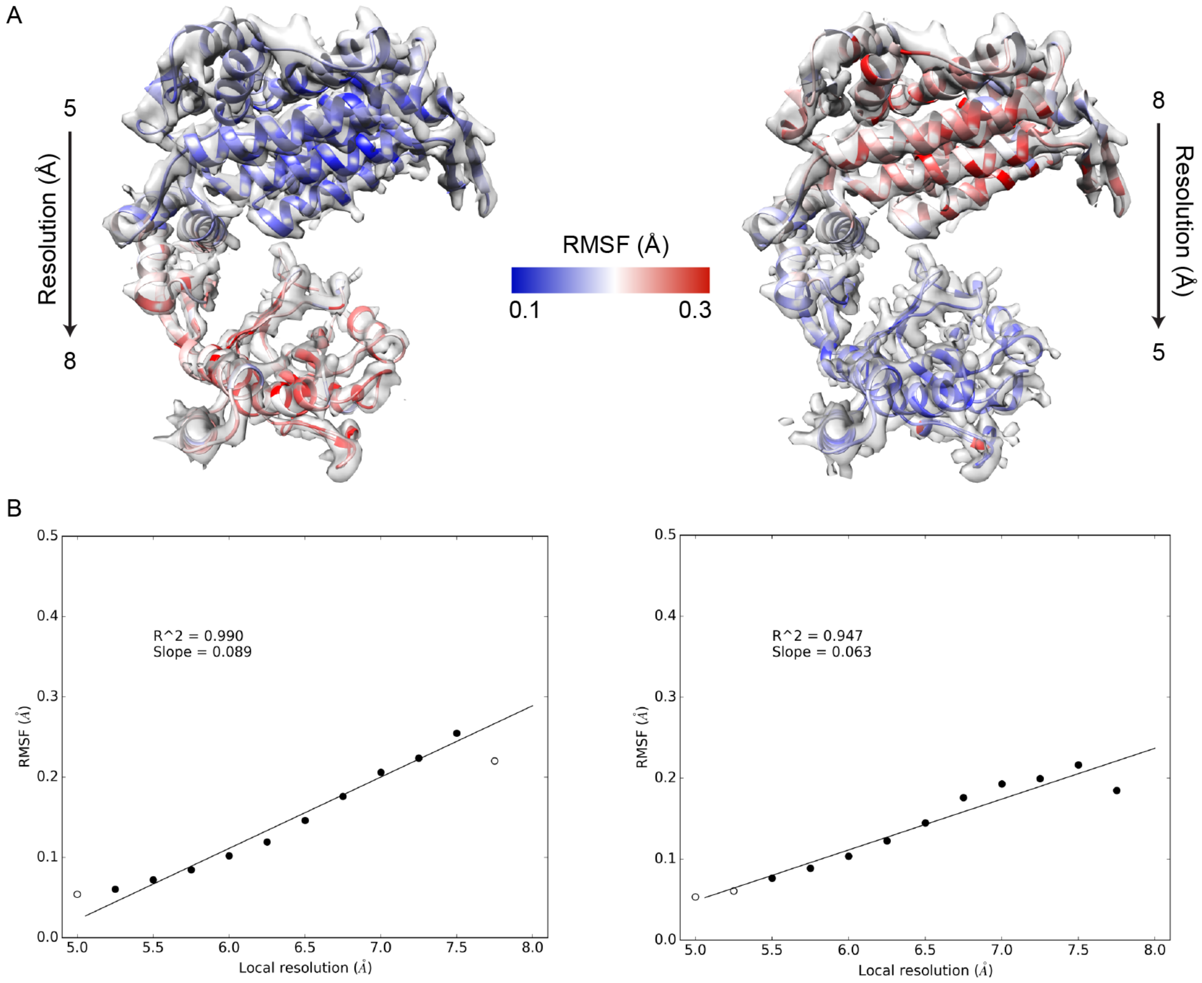


Figure 2: Effect of local resolution on fluctuations. (A) Left half, simulated map with 5Å resolution at the top and 8Å resolution at the bottom, with continuously interpolated resolutions in between. Right side is the same analysis on map with reversed resolutions. Structures are colored by the root mean square fluctuations of the Cα position during sampling. (B) Same arrangement as (A), plotting root mean squared fluctuation as a function of the average local resolution of each atom. Circles are averages at each resolution. Empty circles had <20 atoms and were not included in fit.

two effects: the scores overall get worse, and the score gradually loses the ability to distinguish models. In contrast, for small values of ξ, the differences between models become very severe, since this is effectively assuming that all voxels are independent. Notably, a minimum is apparent for each structure, and the best aligned structure has the absolute minimum at the lowest value of ξ.

The previous test showed that higher values of ξ increase tolerance, as expected. We next wanted to test whether a similar effect was observed between data sets of the same system at multiple resolutions. We expected that, as before, lower resolution data should have more tolerance. Continuing to study the behavior of the score for γ-secretase, we scored the decoy set (described in **Figure 3**) against two different maps, 3.4Å (EMDB:3061, Bai et al. 2015) and 5.4Å (EMDB: 2678) (**Figure 4**). Notably, the score is much sharper for the high resolution map, and has a smaller range of convergence (as visible by the score being roughly flat for the 5.4Å map under 1.5Å, although this is partially an artifact due to the RMSD comparison being to the deposited structure for the high-resolution map, which does not fit perfectly into the low-resolution map). Furthermore, the relationship between local resolution and local fluctuations is still apparent (even though the 3.4Å map is much more uniform in resolution).

Note that for the tests with γ-secretase at 3.4Å and 5.4Å we fixed ξ to 1.15 and 0.8, respectively. These numbers were chosen by finding the value of ξ for the most accurate structure that optimizes the score (*i.e.*, looking at plots like **Figure 3** and finding the lowest value for the "RMSD 0.0" structure). Surprisingly, the slopes of the two fluctuation plots are quite different: 0.516 for the 3.4Å map, and 0.225 for the 3.4Å map, and both are much higher than the slopes in **Figure 2**. This may be a consequence of the choices for ξ, or perhaps the effect of local resolution on RMSF is nonlinear; more effort is needed to understand this effect.

Sampling the model posterior leads to more accurate ensembles

Finally we explored the accuracy of the model ensemble produced during a hard sampling problem, fitting Trpv1 (EMDB: 5778, Liao et al. 2013) with molecular dynamics (**Figure 5**). The best way

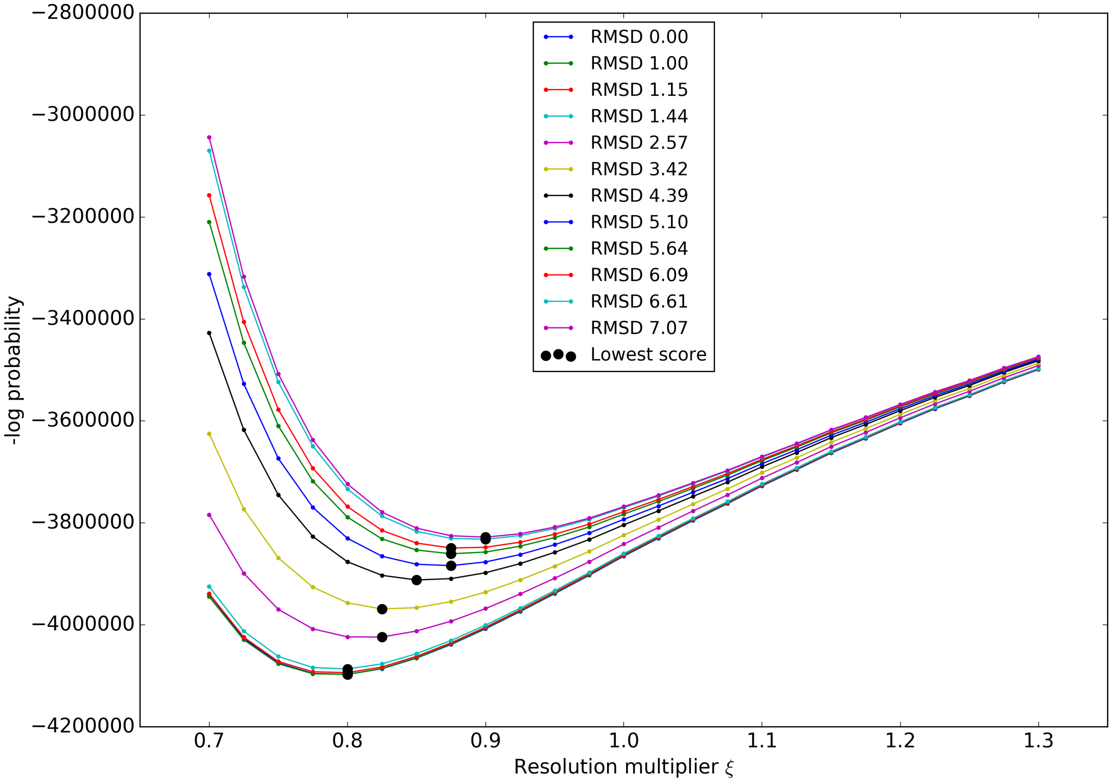


Figure 3: Effect of setting the correlation to various constant multiples of the local resolution. Target system is γ-secretase at 5.4Å resolution. Each line corresponds to a decoy of the low-RMSD structure. The decoy models were generated with a short MD simulation using only a stereochemistry force field with secondary structure restraints. Black dots are the best-scoring values of ξ for each model.

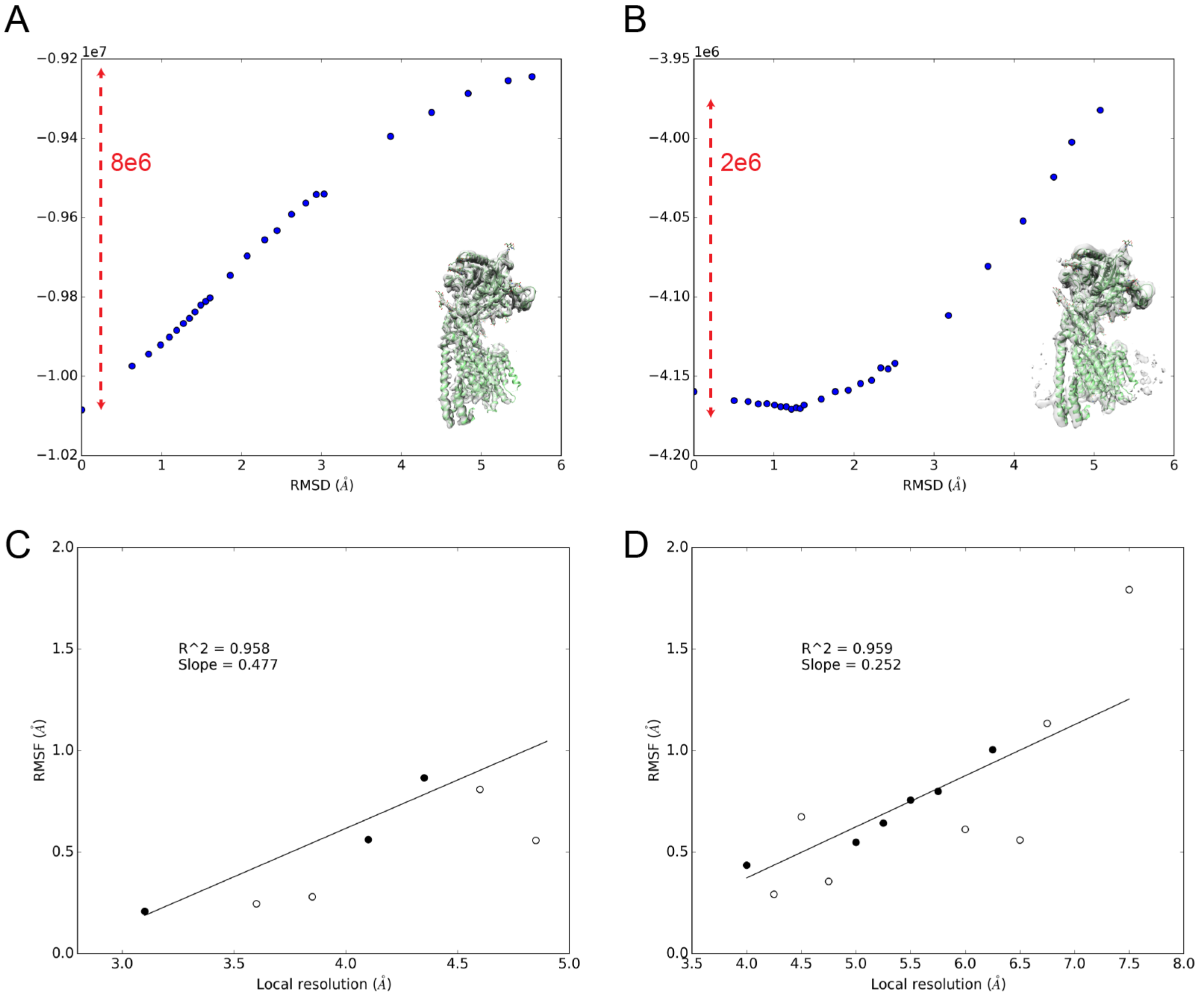


Figure 4: Scoring the same system at multiple resolutions. (A) Fitting scores for γ-secretase at 3.4Å. We scored the same decoy set described in Figure 3 against the map to observe the effect of resolution on the score. X-axis is the root mean squared deviations between each decoy model and the deposited model for γ-secretase at 3.4Å. Red dotted line and number indicate the spread between best and worst scoring model in the set. (B) Decoy set fitted against γ-secretase at 5.4Å. Note that this is not the same test as Figure 3, which varied the resolution multiplier ξ; here, we have fixed the resolution multiplier at the lowest-scoring value for each structure, and varied the map. (C) Root mean squared fluctuations as a function of local resolution for γ-secretase at 3.4Å and (D) γ-secretase at 5.4Å.

to assess accuracy of the ensemble would be to use a higher-resolution reference; however none is available in this case. The accuracy of the ensemble as a whole, though, can be evaluated by checking if the variety of models visited during sampling matches the uncertainty in the map. We expected that averaging the electron density of the model ensemble should produce a density map that looks like the target map—this would indicate complete sampling took place.

We began with the deposited model (PDB: 3j9j), randomizing it with a short MD simulation, with a standard CHARMM force field and secondary structure restraints, up to an RMSD of 5Å. We then sampled models using the model posterior and, separately, with a traditional cross-correlation based score with weight set to 10^5 (chosen by guessing and checking if the model fit well). Once again the fluctuations are highly correlated with the local resolution. To evaluate ensemble accuracy, we created an "ensemble map" by simulating density maps (at Nyquist resolution, or two times the voxel size) of each model in the ensemble of solutions after convergence and then averaged all the maps together, for each score separately. We compared each ensemble map against the target map and found that the cross-correlation was higher for the new probabilistic score (0.6576) than for the conventional score (0.6189), indicating a better capture of the original data.

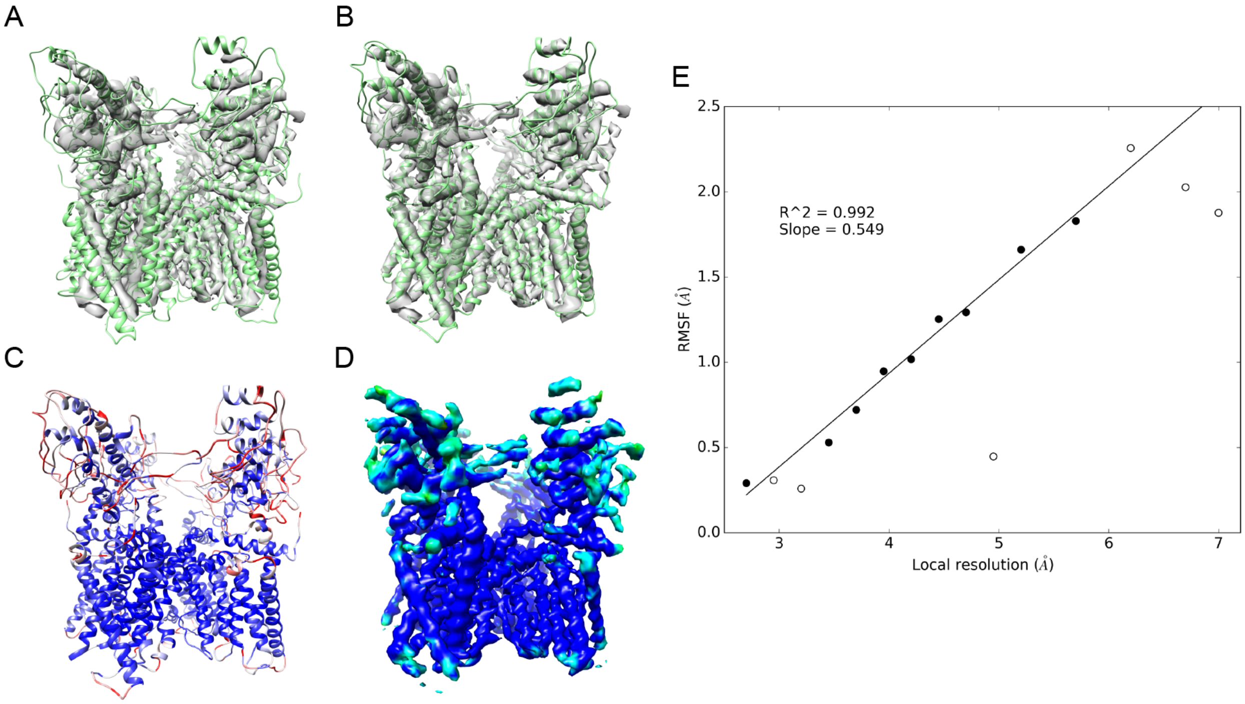


Figure 5: Sampling structures of Trpv1 using the probabilistic score. (A) Initial structure used for fitting. (B) Top-scoring fit using the probabilistic score. (C) Same structure as in (B) but colored by root mean square fluctuations of Cα. (D) Local resolution map of Trpv1. (E) Local fluctuations plotted against local resolution for this fit.

Discussion

We have implemented, for the first time, a probabilistic scoring function for single-particle cryo-electron microscopy data that directly models the varying uncertainty of the data points and the correlations between the neighboring data points. The benefit of this approach is that the probabilistic score is automatically weighted by the uncertainty of the data, as estimated by its local resolution. As a result, the score has few arbitrary parameters, only requiring the user to determine the approximate total mass of the input EM density map.

We have shown that the local resolution directly controls the fluctuations of the model during sampling against simulated data (**Figure 2**), indicating that the variable fluctuations in real data are at least partially data-driven, and not solely due to inherent flexibility of the system. Moreover, the tolerance of the scoring function for violations of the data is set by the local resolution (**Figure 3**). For real data, γ-secretase at 3.4Å and 5.4Å, we found that the score range between good and bad models was much wider in the higher quality data set, indicating that the probabilistic model is more tolerant at lower resolutions (**Figure 4**). This effect normally requires user intervention, but here it is a direct result of setting the uncertainty based on the local resolution. Finally, we showed that for a hard sampling problem, Trpv1 at 3.2Å, it was possible to sample good structures from a starting configuration with high initial deviation (**Figure 5**). Crucially, we found that the ensemble of sampled structures better resembled the target density map than the ensemble from a traditional scoring technique.

These results show that the score is utilizing the local resolution to guide the fit, and that the approach sets the weight automatically based on data quality. However, more work is needed to prove that the accurate ensemble is being generated during sampling. In the future, we will model into EM density maps that have a high-resolution reference, such as a higher-quality EM density map or an X-ray crystal structure. A critical barrier to such testing is ensuring that reference structure is sufficiently close to the structure from which the fitted EM density map was derived (on the scale of the uncertainty of the fitted EM density map). For example, in the γ-secretase comparison in **Figure 4**, the best-scoring solution does not have the lowest RMSD; this finding is likely due to the maps being slightly different (such as a systematic error in the lower-resolution map).

Additionally, we will test the effect of allowing the resolution multiplier ξ to vary simultaneously with sampling the model coordinates. Such a variation likely increases the width of the ensemble, as can be seen in **Figure 3**; at larger correlation multiples, the scoring differences between structure is smaller, so we can expect higher fluctuations farther from the model. Hybrid sampling may also improve the radius of convergence by encouraging more motion when far from the correct solution. We also will explore fitting the local resolution map during modeling (*e.g.*, using N resolution "knots" at locations in the density and sampling their local correlations). A large computational burden can be avoided by using suitable priors to keep the correlations smooth. Extensions of the SPDE allowing this effect are discussed in (Fuglstad et al. 2015).

The new probabilistic score is implemented in our open source *Integrative Modeling Platform* (IMP) package (http://integrativemodeling.org) (Russel et al. 2012), which will allow it to be used simultaneously with many other scoring functions. When other data are present, the benefit of using probabilistic scores is apparent: no expertise in choosing arbitrary weights is needed by the user.

Addendum 1: Non-stationary precision matrix

The first goal is to solve the non-stationary SPDE:

$$(\kappa(\mathbf{s})^2 - \Delta)(u(\mathbf{s})) = \mathscr{W}(\mathbf{s}) \tag{2.13}$$

where $\kappa(s)$ is related to the locally-varying range, $\Delta = \nabla \cdot \nabla$ is the Laplacian (sum of second derivatives in each dimension, or the divergence of the gradient), and $\mathscr{W}$ is a white noise process. Following the notation in Fuglstad et al. 2013, we will solve this differential equation with the finite element method, where each element is a single voxel of the EM density map. Suppose the size of the map is $M \times N \times P$ voxels, and each voxel is symmetric with edge length h and volume $V = h^3$. Furthermore assume periodic boundary conditions, which makes the calculation much simpler and only requires padding the input density map. Consider integrating over one voxel $E_{i,j,k}$:

$$\int_{E_{i,j,k}} \kappa(\mathbf{s})^2 u(\mathbf{s})\, d\mathbf{s} - \int_{E_{i,j,k}} \Delta u(\mathbf{s})\, d\mathbf{s} = \int_{E_{i,j,k}} \mathscr{W}(\mathbf{s})\, d\mathbf{s} \tag{2.14}$$

In the following, we calculate values for each term in Equation 2.14 and solve for the distribution of u. We can approximate the first term on the left by assuming that the range and the density $u(\mathbf{s})$ are constant within a voxel:

$$\int_{E_{i,j,k}} \kappa(\mathbf{s})^2 u(\mathbf{s}) dV \approx V \kappa_{i,j,k}^2 u_{i,j,k} \tag{2.15}$$

Thus, considering $u(\mathbf{s})$ as a vector of density values on each voxel, we can write the first term in Equation 2.13 as a matrix-vector product:

$$\kappa(\mathbf{s})^2 u(\mathbf{s}) = V D_{\kappa^2} \mathbf{u} \tag{2.16}$$

where D_{κ^2} is the diagonal matrix whose diagonal entries are $\kappa_{i,j,k}^2$.

Now consider the second term on the left in Equation 2.14. We can simplify this volume integral using the divergence theorem:

$$\int_{E_{i,j,k}} \Delta u(\mathbf{s})\,d\mathbf{s} = \int_{E_{i,j,k}} \nabla\cdot\nabla u(\mathbf{s})\,d\mathbf{s} = \oint_{\partial E_{i,j,k}} \nabla u(\mathbf{s})\cdot\mathbf{n}\,d\sigma \tag{2.17}$$

where ∂ indicates we are calculating the surface integral over the boundary of the voxel, and $\mathbf{n}$ is the normal vector to that surface.

The boundary of the voxel consists of six faces, each aligned along the x, y, or z axis. The gradient of u is given by $\nabla u(\mathbf{s}) = \frac{\partial u(\mathbf{s})}{\partial x}\hat{x} + \frac{\partial u(\mathbf{s})}{\partial y}\hat{y} + \frac{\partial u(\mathbf{s})}{\partial z}\hat{z}$. Thus, for each face $\nabla u(\mathbf{s})\cdot\mathbf{n}$ only has one non-zero component, which is the derivative in the direction orthogonal to that face. For example let us solve the face pointing to the positive x axis:

$$\begin{aligned}\oint_{\sigma^{x+}_{i,j,k}} \nabla u(\mathbf{s})\cdot\mathbf{n}\,d\sigma &= \oint_{\sigma^{x+}_{i,j,k}} \frac{\partial u(\mathbf{s})}{\partial x}\,d\sigma \\ &\approx h^2 \frac{\partial u(\mathbf{s})}{\partial x}\Big|_{u_{i+1/2,j,k}} \\ &\approx h^2\left(\frac{u_{i+1,j,k}-u_{i,j,k}}{h}\right)\end{aligned} \tag{2.18}$$

where in the second line we assume the derivative at all points on the face are equal to the derivative at the center of the face, and in the third line we approximate the derivative as the average change from the neighboring voxel to the current one.

For the face pointing to the negative x-axis, the calculation is nearly the same:

$$\begin{aligned}\oint_{\sigma^{x-}_{i,j,k}} \nabla u(\mathbf{s})\cdot\mathbf{n}\,d\sigma &= -\oint_{\sigma^{x-}_{i,j,k}} \frac{\partial u(\mathbf{s})}{\partial x}\,d\sigma \\ &\approx -h^2 \frac{\partial u(\mathbf{s})}{\partial x}\Big|_{u_{i-1/2,j,k}} \\ &\approx -h^2\left(\frac{u_{i,j,k}-u_{i-1,j,k}}{h}\right)\end{aligned} \tag{2.19}$$

Repeating these approximations for all faces we get:

$$\begin{aligned}\oint_{\partial E_{i,j,k}} \nabla u(\mathbf{s}) \cdot \mathbf{n}\, d\sigma \approx \\ h\left[(u_{i+1,j,k} + u_{i-1,j,k} + u_{i,j+1,k} + u_{i,j-1,k} + u_{i,j,k+1} + u_{i,j,k-1}) - 6u_{i,j,k}\right]\end{aligned} \tag{2.20}$$

We have thus approximated the integral of the divergence of $u(\mathbf{s})$ with averages of neighboring values. When we use this approximation for all the voxels, we can store these relationships in a matrix:

$$\int_{E_{i,j,k}} \kappa(\mathbf{s})^2 u(\mathbf{s}) dV \approx \mathbf{A_I u} \tag{2.21}$$

where $\mathbf{A_I}$ has size $MNP \times MNP$. To simplify notation, define i_p and i_n to be the rows (in the original density map) of voxels before and after row i. Similarly define j_p, j_n, k_p, and k_n. Note that due to periodic boundary conditions, these values wrap around the density map (*e.g.* if i is $M-1$ then i_n is 0. With this notation we write the entries of one row $\mathbf{A_I}$ corresponding to voxel $u_{i,j,k}$:

$$\begin{aligned}
(\mathbf{A_I})_{kMN+jM+i,kMN+jM+i} &= -6h \\
(\mathbf{A_I})_{kMN+jM+i,k_pMN+jM+i} &= h \\
(\mathbf{A_I})_{kMN+jM+i,k_nMN+jM+i} &= h \\
(\mathbf{A_I})_{kMN+jM+i,kMN+j_pM+i} &= h \\
(\mathbf{A_I})_{kMN+jM+i,kMN+j_nM+i} &= h \\
(\mathbf{A_I})_{kMN+jM+i,kMN+jM+i_p} &= h \\
(\mathbf{A_I})_{kMN+jM+i,kMN+jM+i_n} &= h
\end{aligned} \tag{2.22}$$

Finally, the right hand side Equation 2.14 is the integral of a white noise process over a finite element with volume V, which gives a Gaussian variable with mean 0 and variance V (Fuglstad et al. 2013). Each of these variables is independent for each voxel since the voxels do not overlap.

Thus:

$$\int_{E_{i,j,k}} \mathscr{W}(\mathbf{s})\, d\mathbf{s} = \sqrt{V} z_{i,j,k} \tag{2.23}$$

where $z_{i,j,k}$ is a standard Gaussian variable for each voxel. Now we can combine Equations 2.16,2.21, and 2.23 together to form the discrete approximation of the full SPDE (Equation 2.13):

$$V\mathbf{D}_{\kappa^2}\mathbf{u} - \mathbf{A_I}\mathbf{u} = \sqrt{V}\mathbf{z} \tag{2.24}$$

where $\mathbf{z} \sim \mathscr{N}_{M,N,P}(\mathbf{0}, \mathbf{I}_{M,N,P})$ is a multivariate normal distribution with identity covariance (uncorrelated variables). Thus we can re-write Equation 2.24 as:

$$\mathbf{z} = V^{-1/2}\mathbf{A}\mathbf{u} \tag{2.25}$$

where $\mathbf{A} := V\mathbf{D}_{\kappa^2} - \mathbf{A_I}$. Following Fuglstad et al. 2013, this gives the joint distribution of $\mathbf{u}$:

$$\begin{aligned}
\pi(\mathbf{u}) &\propto \pi(\mathbf{z}) \propto \exp\left(-\frac{1}{2}\mathbf{z}^T\mathbf{z}\right) \\
\pi(\mathbf{u}) &\propto \exp\left(-\frac{1}{2V}\mathbf{u}^T\mathbf{A}^T\mathbf{A}\mathbf{u}\right) \\
\pi(\mathbf{u}) &\propto \exp\left(-\frac{1}{2}\mathbf{u}^T\mathbf{Q}\mathbf{u}\right)
\end{aligned} \tag{2.26}$$

where $\mathbf{Q} = \frac{1}{V}\mathbf{A}^T\mathbf{A}$ is the precision matrix of u. This is a sparse matrix with a maximum of 37 non-zero elements on each row, consisting of the point itself, its 6 closest neighbors, and the 6 closest neighbors of those neighbors.

Addendum 2: Derivatives

This is the likelihood:

$$p(y|V) = (2\pi)^{-\frac{Nvox}{2}} |K|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(y-V)^{\intercal}K^{-1}(y-V)\right) \tag{2.27}$$

Where the contribution to each voxel i from atom j is given by:

$$V_{ij} = \frac{Z_j}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{\vec{r}_j - \vec{r}_i}{\sigma}\right)^2\right] \tag{2.28}$$

Now due to the product rule:

$$\begin{aligned} \frac{dp(y)}{d\vec{r}_j} &= -\frac{1}{2}p(y)\frac{d}{d\vec{r}_j}(y-V)^{\intercal}K^{-1}(y-V) \\ &= p(y)(y-V)^{\intercal}K^{-1}\frac{d(V)}{d\vec{r}_j} \end{aligned} \tag{2.29}$$

Where each component i of the derivative is a vector given by:

$$\frac{dV_i}{d\vec{r}_j} = V_{ij}\left(\frac{\vec{r}_i - \vec{r}_j}{\sigma^2}\right) \tag{2.30}$$

References

Bai, X.-C. et al. 2015. “Sampling the conformational space of the catalytic subunit of human γ-secretase.” *Elife* 4:e11182.

Eidsvik, J. et al. 2014. “Estimation and prediction in spatial models with block composite likelihoods.” *Journal of Computational and Graphical Statistics* 23 (2): 295–315.

Fuglstad, G.-A. et al. 2015. “Constructing Priors that Penalize the Complexity of Gaussian Random Fields.” *arXiv preprint arXiv:1503.00256.*

Fuglstad, G.-A. et al. 2013. “Exploring a new class of non-stationary spatial Gaussian random fields with varying local anisotropy.” *arXiv preprint arXiv:1304.6949.*

Gilboa, E., Saatci, Y., and Cunningham, J. P. 2015. “Scaling Multidimensional Inference for Structured Gaussian Processes.” *IEEE Trans Pattern Anal Mach Intell* 37 (2): 424–36.

Habeck, M., Rieping, W., and Nilges, M. 2006. “Weighting of experimental evidence in macromolecular structure determination.” *Proceedings of the National Academy of Sciences of the United States of America* 103 (6): 1756–1761.

Han, I., Malioutov, D., and Shin, J. 2015. “Large-scale log-determinant computation through stochastic Chebyshev expansions.” *arXiv preprint arXiv:1503.06394.*

Higdon, D., Swall, J., and Kern, J. 1999. “Non-stationary spatial modeling.” *Bayesian Statistics 6*: 761–768.

Kaufman, C. G., Schervish, M. J., and Nychka, D. W. 2008. “Covariance tapering for likelihood-based estimation in large spatial data sets.” *Journal of the American Statistical Association* 103 (484): 1545–1555.

Kucukelbir, A., Sigworth, F. J., and Tagare, H. D. 2014. “Quantifying the local resolution of cryo-EM density maps.” *Nat Methods* 11 (1): 63–5.

Liao, M. et al. 2013. “Structure of the TRPV1 ion channel determined by electron cryo-microscopy.” *Nature* 504 (7478): 107–112.

Lindgren, F., Rue, H., and Lindström, J. 2011. "An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 73 (4): 423–498.

Lu, P. et al. 2014. "Three-dimensional structure of human gamma-secretase." *Nature* 512 (7513): 166–170.

Malioutov, D. M., Johnson, J. K., and Willsky, A. S. "Low-rank variance estimation in large-scale GMRF models." In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, 3:III–III. IEEE.

Paciorek, C. J. and Schervish, M. J. 2006. "Spatial Modelling Using a New Class of Nonstationary Covariance Functions." *Environmetrics* 17 (5): 483–506.

Rasmussen, C. E. and Williams, C. K. I. 2006. *Gaussian processes for machine learning*. Cambridge, Mass. ; London: MIT.

Rieping, W., Habeck, M., and Nilges, M. 2005. "Inferential structure determination." *Science* 309 (5732): 303–6.

Rue, H. and Held, L. 2005. *Gaussian Markov random fields: theory and applications*. CRC Press.

Russel, D. et al. 2012. "Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies." *PLoS Biol* 10 (1): e1001244.

Scheres, S. H. W. 2012. "A Bayesian view on cryo-EM structure determination." *J Mol Biol* 415 (2): 406–18.

Simpson, D., Lindgren, F., and Rue, H. 2012. "Think continuous: Markovian Gaussian models in spatial statistics." *Spatial Statistics* 1:16–29.

Topf, M. et al. 2008. "Protein structure fitting and refinement guided by cryo-EM density." *Structure* 16 (2): 295–307.

Ward, A. B., Sali, A., and Wilson, I. A. 2013. "Integrative structural biology." *Science* 339 (6122): 913–915.

Whittle, P. 1954. “On Stationary Processes in the Plane.” *Biometrika* 41 (3-4): 434–449.

Chapter 3

Structure of γ-tubulin small complex based on a cryo-EM map, chemical cross-links, and a remotely related structure

Abstract

Modeling protein complex structures based on distantly related homologues can be challenging due to poor sequence and structure conservation. Therefore, utilizing even low-resolution experimental data can significantly increase model precision and accuracy. Here, we present models of the two key functional states of the yeast γ-tubulin small complex (γTuSC): one for the low-activity "open" state and another for the higher-activity "closed" state. Both models were computed based on remotely related template structures and cryo-EM density maps at the 6.9 Å and 8.0 Å resolution, respectively. For each state, extensive sampling of alignments and conformations was guided by the fit to the corresponding cryo-EM density map. The resulting good-scoring models formed a tightly clustered ensemble of conformations in most regions. We found significant structural differences between the two states, primarily in the γ-tubulin subunit regions where the microtubule binds. We also report a set of chemical cross-links that were found to be consistent with equilibrium between the open and closed states. The protocols developed here have been incorporated into our open-source Integrative Modeling Platform (IMP) software package (http://integrativemodeling.org), and can therefore be applied to many other systems.

Introduction

Biologists are often interested in computing structural models of protein assemblies for which electron microscopy density maps and atomic structures of remotely related systems are available. In such cases, the integrative approach that relies on input structural information of multiple types can provide models that are more accurate, precise, and complete than models based on traditional methods, such as X-ray crystallography, NMR spectroscopy, and electron microscopy (Alber et al. 2008; Alber et al. 2007; Russel et al. 2012; Sali et al. 2015; Ward et al. 2013). Key challenges include: data ambiguity, such as regions of a density map that are not assignable to particular components of the system or cross-links that may apply to one of many states or copies of the structure; information uncertainty, including the limited resolution of the map, errors in the template structures, and target-template differences; and data completeness, including missing regions of the density map and incomplete coverage by the templates and cross-links (Schneidman-Duhovny et al. 2014). While numerous tools exist to construct models based on density maps (Topf et al. 2008; Trabuco et al. 2008), it is essential to find all models that fit the data, allowing an estimate of model precision. Here, we describe an *ad hoc* approach to inferring a pseudo-atomic model from a cryo-EM density map and atomic structures of related proteins, with an estimate of model precision based on variation among good-scoring models, and an estimate of model accuracy based on agreement with an independently determined set of chemical cross-links. We illustrate the approach by its application to the γ-tubulin small complex (γTuSC).

γ-tubulin complexes control the location and timing of microtubule nucleation. γTuSC is a 300 kDa complex consisting of four components: GCP2, GCP3, and two copies of γ–tubulin in a V-shaped structure with the γ–tubulin subunits at the top (Choy et al. 2009; Kollman et al. 2008). As revealed by an 8 Å cryo-EM map, the complex assembles into a single helical turn in yeast consisting of 7 γTuSC units, leaving 13 exposed γ-tubulins, thus allowing 13-protofilament microtubule filaments to template from the complex (Kollman et al. 2010). γTuSC comprises the "minimal" nucleation machinery, as it lacks numerous accessory proteins found in related complexes from other species. In previous studies, we have shown that γTuSC adopts two conformations that differ significantly in the geometry of the γ-tubulin ring—a low-activity "open" state and a higher-activity "closed" state—and have proposed that this conformational switch provides a mechanism for regulating microtubule nucleation *in vivo* (Kollman et al. 2011). We have reported the cryo-EM structure of the "open" state at 8 Å resolution, and more recently reported the cryo-EM structure of the "closed" state at 6.9 Å resolution, by trapping this transient conformation with engineered disulfides (Kollman et al. 2015). Here, we describe reliable pseudo-atomic models of both open and closed complexes, and present new data consistent with equilibrium between two conformational states.

Approach

Closed-state model building

We followed an iterative procedure (**Figure 1**) to generate and fit comparative models (Topf et al. 2006). The initial alignment of the entire TUBGCP family was performed in Promals3D (Pei et al. 2008). Additionally, we aligned yeast γ-tubulin to representative proteins from the tubulin family (including human γ-tubulin, the template). After alignment, regions of the human GCP4 sequence that were not present in the crystal structure were omitted. We utilized an iterative procedure to improve the alignment and optimize models with respect to the density map, including the following steps:

1. Four copies of each template (human GCP4 and human γ-tubulin) were rigidly docked into the closed-state map to form an approximate shape of two side-by-side γTuSC structures using UCSF Chimera (Pettersen et al. 2004).

2. The template complex in step 1 was used as the basis for simultaneous homology modeling of yeast GCP2, GCP3, and γ-tubulin in MODELLER using the current alignment. Symmetry restraints were added to preserve the complex structure.

3. The γTuSC homology model was flexibly fitted into the closed-state density map using

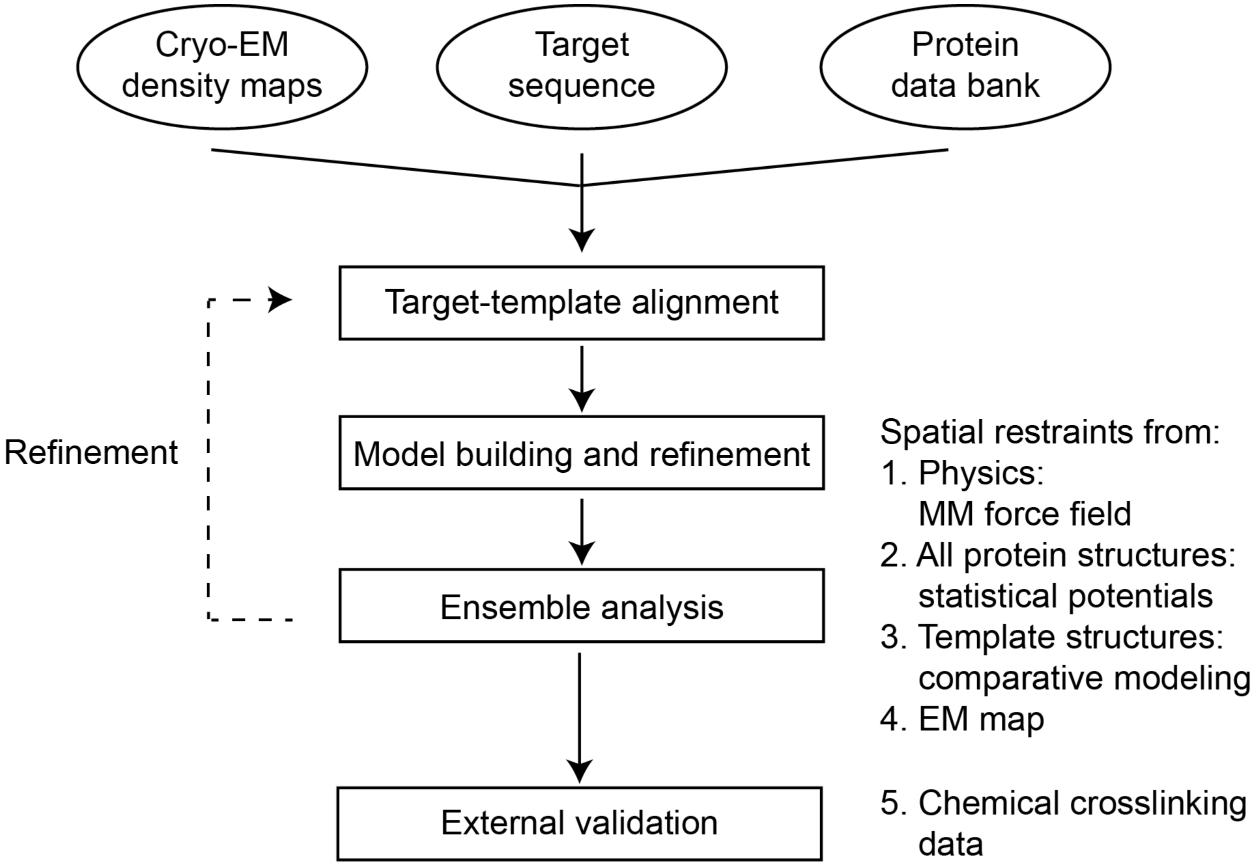

Cryo-EM density maps
Target sequence
Protein data bank
Target-template alignment
Refinement
Model building and refinement
Ensemble analysis
External validation
Spatial restraints from:
1. Physics: MM force field
2. All protein structures: statistical potentials
3. Template structures: comparative modeling
4. EM map
5. Chemical crosslinking data

Figure 1: Method overview.

Molecular Dynamics Flexible Fitting (MDFF) (Trabuco et al. 2008), with additional secondary structure restraints. Symmetry restraints were added between the two copies of γTuSC. The “gscale” parameter of MDFF was set to 1.0, and 200ps simulations were run.

4. Upon inspection of the fitted model, the pairwise human-GCP4/yeast-GCP2 and human-GCP4/yeast-GCP3 alignments were edited. Changes included adding secondary structure restraints when these elements were clearly observed in the density map and were predicted using PSIPRED (Buchan et al. 2013) and removing long insertions (>5 residues) unless the insertions could be unambiguously assigned to density. The γ-tubulin alignment was left unchanged.

Steps 2-4 were repeated until no further adjustments to the alignment improved the model. With the alignment fixed via the above procedure, we produced 300 comparative models as in Step 2. Finally, these models were optimized into the closed-state density map using MDFF with the same restraints as above. Results contained a number of clear outliers, with structural helices crossing obvious helices in the map. We removed these common local minima by performing k-means clustering with k=10, keeping the largest two clusters containing 158 total structures, which formed a narrow ensemble with average Cα root-mean-square fluctuation (RMSF) 1.88 Å. The 5 top scoring structures from each cluster were deposited into the PDB: 5FLZ.

Open-state model building

Reasoning that the closed-state model was more accurate than the starting comparative model, we used the closed-state model ensemble as the initial structures for open-state modeling. For every closed model, each component was rigidly fitted into the open-state density map. These models were refined into the open-state density map using MDFF, with the “gscale” parameter set to 0.1 (reflecting our increased uncertainty in the data due to its lower resolution) and allowed to sample for 250ps. Secondary structure, domain (rigidity), and symmetry restraints were added as in the closed state. The open-state ensemble consisted of all fitted structures, since the lower resolution did not allow us to remove obvious outliers. We performed k-means clustering with k=10 and selected the top-scoring structure in each cluster—this forms the ensemble deposited into the PDB: 5FM1.

Model evaluation

To evaluate model precision, we calculated the RMSF for each Cα position in the model ensemble. To evaluate sampling density, we divided the ensemble into half-sets, computed the RMSF values for each set, and verified that they were similar. For contact evaluation (**Table 1**), UCSF Chimera's "find H-bond" tool was used on each model in the ensemble, relaxing H-bond parameters by 2 Å and 90°. The final reported contacts were those present in at least 25% of the structures in each ensemble. To evaluate model accuracy, we compared models against external data not used in the modeling process. A set of 135 chemical cross-links, with DSS as the linker, was obtained. We computed the distance between cross-linked residues within the closed- and open-state models, allowing for ambiguity in the cross-links' assignments due to the presence of two γ-tubulin molecules in γTuSC, as well as multiple copies of γTuSC. A cross-link was considered a "violation" if the median ensemble distance was greater than a threshold in both the closed- and open-state models. The maximal cross-link distance of 35 Å was based on flexibility of the cross-linker (Chen et al. 2010).

Cross-linking of recombinant γTuSC and mass spectrometry analysis

γTuSC (146 Mg protein in 331 ML 40 mM HEPES, 100 mM NaCl, pH 8) was cross-linked for 2 min at room temperature with disuccinimidyl suberate (Pierce, 0.86 mM final). The reaction mix was quenched with 26 ML of 500 mM NH4HCO3 and the buffer was exchanged to 40 mM HEPES, 500 mM NaCl, pH 7.5 using protein desalting spin columns (Pierce) according to the manufacturer's instructions. Two 90 uL aliquots of cross-linked protein were subsequently reduced with 10 mM dithiothreitol for 30 min at 37oC and alkylated with 15 mM iodoacetamide for 30 min at room temperature. Heavy oxygen labeling (Zelter et al. 2010) was performed by adding 25%

volume of heavy water to one (labeled) aliquot. A second, unlabeled, sample was produced by adding 25% volume of standard water. Both aliquots were then separately subjected to overnight digestion with trypsin at a substrate to enzyme ratio of 60:1. Samples were acidified with 5 M HCl and stored at -80oC.

0.5 to 1.5 Mg of each sample was loaded onto a fused-silica capillary tip column (75-Mm i.d.) packed with 40 cm of Reprosil-Pur C18-AQ (3-Mm bead diameter, Dr. Maisch). Peptides were eluted from the column at 250 nL/min using a gradient of 2-35% acetonitrile (in 0.1% formic acid) over 120 min, followed by 35-60% acetonitrile over 10 min. Mass spectrometry was performed on a Q-Exactive (Thermo Scientific), operated using data dependent acquisition where a maximum of six MS/MS spectra were acquired per MS spectrum (scan range of m/z 400-1600). At m/z 200, the resolution for MS and MS/MS was 70,000 and 35,000, respectively. Six technical replicates were performed using the heavy oxygen labeled sample and 14 technical replicates were performed using the unlabeled sample. Cross-linked peptides were identified using the Kojak cross-link identification software (Hoopmann et al. 2015) (version 1.4.1) available at (http://www.kojak-ms.org/). The Kojak results of all 20 LCMS runs were combined and exported to Percolator (Kall et al. 2007) to produce a statistically validated set of cross-linked peptide identifications at a false discovery rate threshold of 1%. The full cross-link dataset is available online at http://proxl.yeastrc.org/proxl/viewProject.do?project_id=15 .

Results

Initial model accuracy and coverage

GCP2 and GCP3 are members of the TUBGCP family, which are named Spc97 and Spc98 in yeast. The TUBGCP family also includes GCP2/3/4/5/6 in humans. The architecture of the complex is shown in **Figure 2A**. A crystal structure of human GCP4 provided a suitable starting point for homology modeling of both GCP2 and GCP3 (Guillet et al., 2011) (**Figure 2B,C**). Additionally, a crystal structure of human γ-tubulin (PDB: 3CB2) served as a suitable template (sequence identity

37%) for yeast γ-tubulin. Modeling challenges included the presence of several large insertions in GCP2 and GCP3, regions of GCP4 that were not observed in the crystallographic density, and low sequence identity to GCP4 (13% and 18% for GCP2 and GCP3, respectively) (**Figure 2B**). We first modeled the closed-state structure due to the higher resolution of the corresponding EM map.

Despite the low sequence identity to the homologue GCP4, the initial model of closed-state TuSC, consisting of flexibly fitted homology models of GCP2 and GCP3 and two identical homology models of yeast γ-tubulin, fit the EM map surprisingly well (**Figure 3**). Most secondary structure elements could be uniquely assigned to regions of the map; however, there were clear errors in the length and location of some helices. We improved the model using an iterative process, editing the GCP2/3/4 alignment to improve the fit and using other sources of information to reduce over-fitting. The best-scoring model resolves the locations of all secondary structure elements, though many loops were difficult to localize **Figure 4A**. Large insertions are still missing from the model; these would likely need higher-resolution data to complete *de novo* building of these sections.

In comparison to the closed-state density map, the open-state map is lower resolution. We reasoned that the final closed-state model is likely closer to the open-state structure than the initial homology model. Therefore, we used closed-state models as starting points for open-state modeling. While the resolution limits our ability to precisely localize secondary structure segments, the overall shape could be determined from the density map. The final model has significantly improved cross-correlation in comparison to the starting closed-state model (**Figure 3**).

Estimating model precision and accuracy

The final structures of the closed and open states were selected based on the cross-correlation coefficient against their respective maps. To use any model judiciously, it is essential to assess its precision and accuracy. Here, we estimate the precision by quantifying the variation in the ensemble of good-scoring solutions. For each state, we defined the precision of each Cα position as the root-mean-square fluctuation (RMSF) from the mean position in the models scoring at least

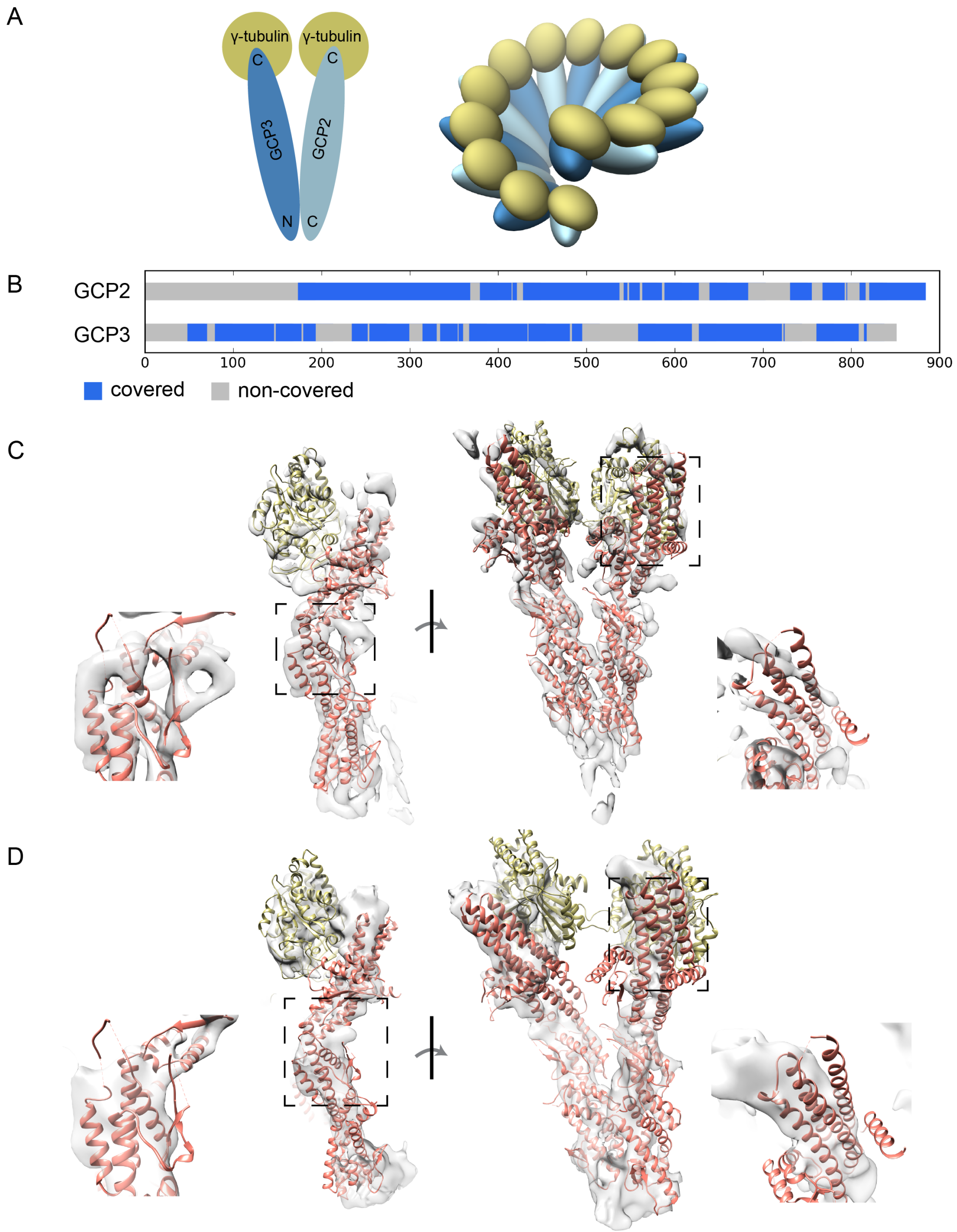


Figure 2: Data overview. (A) Architecture of γ-TuSC. Left, single γ-TuSC V structure with sequence endpoints. These colors are used in all figures. Right, ring structure. (B) Coverage map. Blue regions have structure coverage from the homologue GCP4. (C) Rigid fit of templates into closed cryo-EM density map. Details indicate regions of significant difference between the model and the map. (D) Rigid fit of templates into open cryo-EM density map.

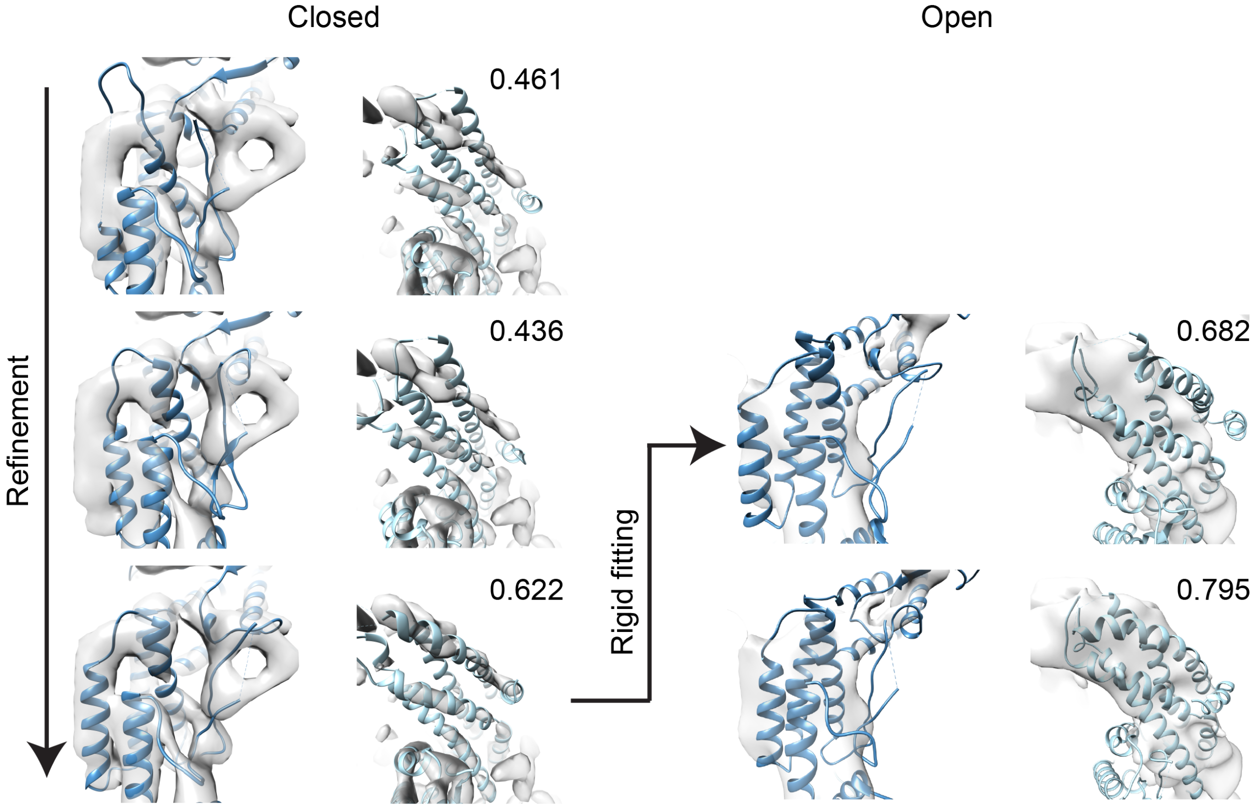


Figure 3: Progress of modeling of each state. On the left, showing three stages of the closed model: initial comparative model, improved alignment model, and flexibly fitted model. Numbers show the cross-correlation coefficient. Arrow to the right indicates using the final structure of the closed state to fit into the open map. Two open-state modeling stages are shown: initial rigid fit and flexible fit.

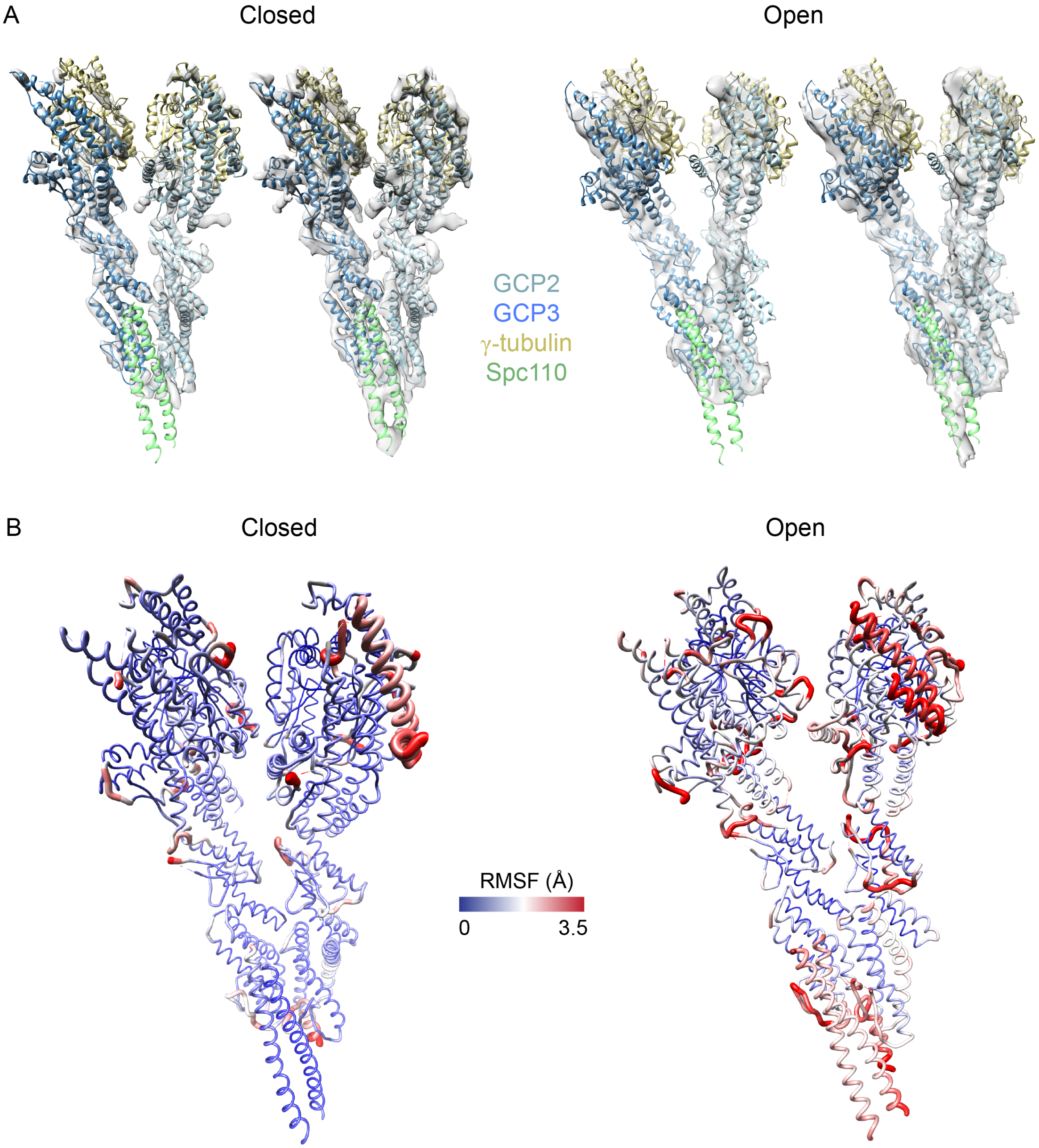


Figure 4: Modeling results. (A) Left, closed-state final fit, shown in closed-state density map at two different contour levels. Right, open-state final fit, again at two contour levels. (B) Sausage plots of the variability in the model ensemble fits. Red regions are more variable than blue regions. Left, closed-state; and right, open-state.

one standard deviation above the median score, with each model fit in the EM map **Figure 4B**. We verified our estimate of precision by showing similar estimates for two random halves of the solution ensemble (data not shown). Regions with high variability were primarily in loops, and appear to be largely a result of the limited map resolution. RMSF is generally larger in the closed state.

Additionally, we validated the γTuSC models by comparing them against a set of DSS cross-links not used in the modeling process (**Figure 5**). An ensemble of models satisfies a given cross-link if the ensemble median of the shortest distance among possible Cα-Cα assignments within a model was less than 35 Å (corresponding to the maximum DSS length with small tolerance (Chen et al. 2010)). Only 67 of the 135 observed cross-links could be used for this assessment because others applied to missing regions. Of these 67 cross-links, 62 were consistent with either the closed or open state, including some that matched exclusively with a single state. Of the 5 "violated" cross-links, one had median distance just over the threshold, and the remaining 4 had median distances over 50Å and are likely false positives.

Comparing the open and closed states

After aligning models from the two states, significant differences—which are larger than the estimated precision—are apparent (**Figure 6**). The differences are particularly notable for the locations of the γ-tubulin subunits, which bind to the minus end α-tubulin subunits within the microtubules. In the open state, microtubule symmetry is broken: GCP2 is bent 8° towards the helical axis and GCP3 is bent 8° backwards, creating varying spacing and orientation between the γ-tubulin subunits. The helical parameters of the open state, rotation of 54.3° and rise of 22.2 Å do not match those of a microtubule, with rotation of 55.4° and rise of 18.7 Å. In contrast, in the closed state, the γ-tubulin subunits are evenly distributed around the ring, with the same side facing the helical axis, matching the symmetry of a microtubule. Helical spacing parameters of the closed state also match those of a microtubule: rotation of 55.4° and rise of 18.8 Å. The arrangement of three contiguous α-tubulin subunits from a mammalian microtubule (PDB: 3JAL) is similar to that of the

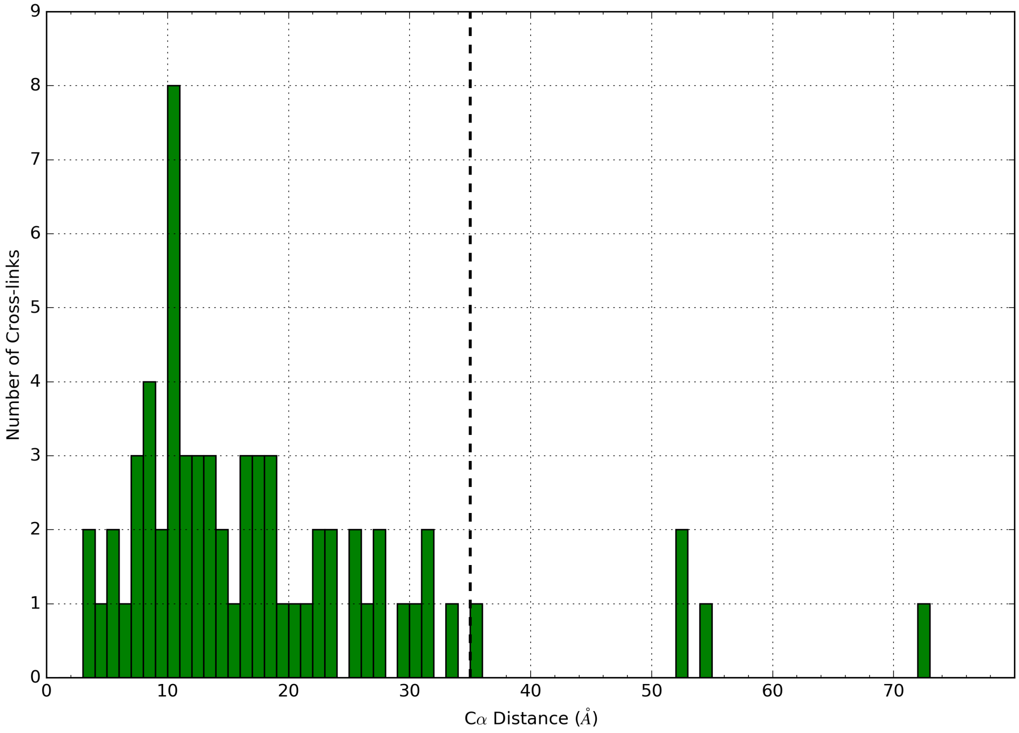


Figure 5: Cross-linking data analysis. Illustrating the histogram of median crosslink distances in the models. Dotted line is "violation" cutoff at 35Å.

yeast γ-tubulin subunits in the closed state, with 90% Cα overlap at 4 Å. While the centers of mass of the γ-tubulin subunits vary significantly between the two states, there is no clear pivot point. The structural changes are widely distributed from the central to C-terminal regions of GCP2 and GCP3.

Two key interfaces underlie the stability of each state: "intra"- TuSC (within the V structure) and "inter"- TuSC (between each copy of the complex). We examined each interface in detail to explore which contacts are unique to each state in our models (**Table 1**). First, we observed that the N-terminal regions of GCP2 and GCP3 are largely static, whereas the C-terminal interfaces shear $\sim$ 15 Å. Secondly, the C-terminal contacts are primarily made between the γ-tubulin subunits. Multiple key interactions at the closed "inter" interface are also preserved at the "intra" interface, supporting the observation of symmetry. Furthermore, many of these contacts also occur at the open "inter" interface, despite the significant conformational differences. This observation is consistent with the apparent rotation around this interface (**Figure 6**). In contrast, the open "intra" interface shares no contacts with the other three, though a similar list of residues are involved in contacts.

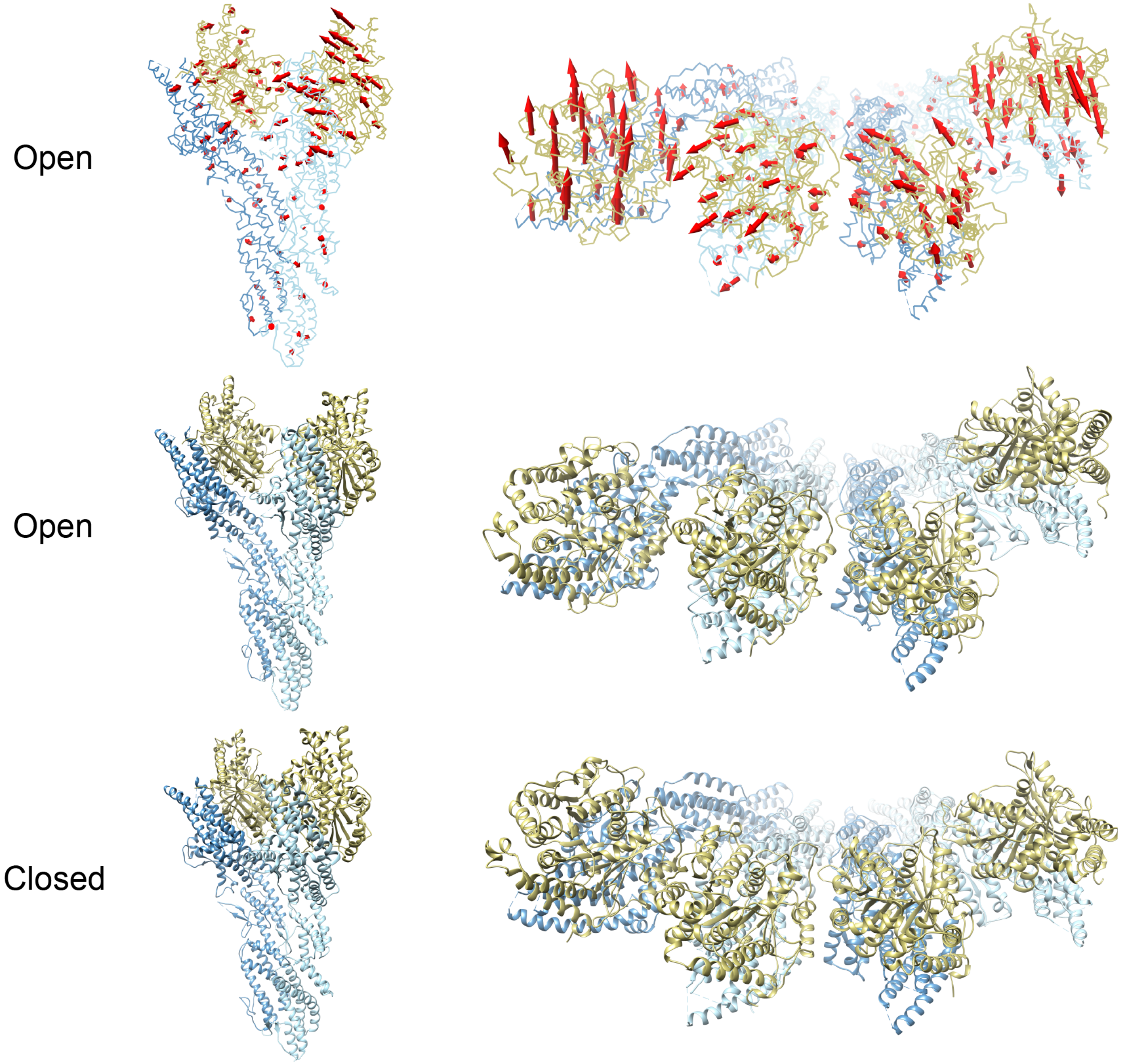


Figure 6: Model comparison. Arrows indicate structural changes from open (top two rows) to closed state (bottom row). Each row depicts only the state listed at the left. Left column is side view of a single γ-TuSC subunit, right column is top view showing two side-by-side subunits.

Contact	Closed/Inter	Closed/Intra	Open/Inter	Open/Intra
55R-291S		X	X	
56E-288G	X	X		
57N-286H	X	X	X	
57N-288G		X	X	
57N-291S			X	
58S-286H	X	X	X	
58S-287K	X	X	X	
58S-289H			X	
58S-327E	X			
58S-370N	X			
59R-282D		X		
59R-283D	X	X		
59R-284I	X	X		
59R-285A			X	
59R-286H	X	X	X	
59R-287K			X	
59R-369E	X			
59R-370N	X		X	
59R-371E	X		X	
90R-291S			X	
90R-300D		X		
121N-218R				X
121N-311N			X	
124D-218R				X
125K-282D				X
125K-283D				X
125K-291S			X	
128D-218R				X
129S-283D				X
129S-290S			X	

Table 1: List of residue pairs involved in key interfaces. "X" indicates that the residue pair is found at that surface.

Discussion

Our goal was to compute and validate models for the two structural states of the γ-tubulin small complex, followed by comparing the two states. For modeling, we used all available information (**Figure 1**). We used the known structures of homologues of GCP2, GCP3, and γ-tubulin to create initial models; used predicted secondary structure segments to explore alternative alignments; and used cryo-EM density maps to refine a model for each state. We also utilized chemical cross-linking data as an external validation, finding that 62/67, or 93%, of the cross-links are consistent with the two models (**Figure 5**). This finding is particularly notable due to the low sequence identity of the templates (13%, 18%, and 37% for GCP2, GCP3, and γ-tubulin, respectively) including large insertions (**Figure 2**). To evaluate the precision of these models, we performed extensive sampling and reported the fluctuations of each Cα atom (**Figure 4B**). As expected, we found that the largest fluctuations were in regions with relatively low density in EM maps, which were typically loop regions. Thus, we are relatively confident in the placement of secondary structure units and the overall fold (**Figure 4A**).

The structural differences between the closed and open states include both rigid body and conformational changes (**Figure 6**). However, despite the significant differences between the states, many key interface contacts are preserved, particularly at the “inter”-γTuSC interface, which may be critical for ring assembly. The “intra”-γTuSC interface, while making a completely different set of contacts between the open and closed states, does involve a similar set of residues, suggesting that any evolutionary pressure that preserves ring assembly may also maintain the activation mechanism. A more complete understanding of the cause of conformational change may require modeling that is more precise, accurate, and complete, including building models for the substantial numbers of insertions (with respect to the template, human GCP4), some of which are located at γTuSC interfaces (**Figure 2**).

The description of the differences between the two end states does not allow us to speculate about the order of events corresponding to a transition between them. However, the cross-linking data suggests that the system exists in equilibrium between the two states, thus transiently exploring both states; the equilibrium between the two states is suggested because both are needed to explain all the cross-links (**Figure 5**). In a previous study (Kollman et al. 2015), we showed that the open state nucleates less well than the disulfide-closed state, suggesting that some activation event likely takes place in the cell to optimize nucleation efficiency.

The approach we have developed here is suitable for challenging problems where structural information is incomplete or low-resolution. The increased uncertainty in these cases requires sampling the full range of models consistent with the available data, evaluating the precision of the ensemble, and, if possible, using independent data to estimate model accuracy. We have incorporated tools to perform these steps in our open-source Integrative Modeling Platform (IMP) package, available at http://integrativemodeling.org (Russel et al. 2012). Specific code used to generate the models is available at https://github.com/integrativemodeling/gamma-tusc. Future improvements include using scoring functions and sampling techniques that do not rely on manually set data weights—for example, Bayesian methods (Rieping et al. 2005).

References

Alber, F. et al. 2007. “Determining the architectures of macromolecular assemblies.” *Nature* 450 (7170): 683–94.

Alber, F. et al. 2008. “Integrating diverse data for structure determination of macromolecular assemblies.” *Annu Rev Biochem* 77:443–77.

Buchan, D. W. et al. 2013. “Scalable web services for the PSIPRED Protein Analysis Workbench.” *Nucleic Acids Res* 41 (Web Server issue): W349–57.

Chen, Z. A. et al. 2010. “Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry.” *EMBO J* 29 (4): 717–26.

Choy, R. M. et al. 2009. “Localization and orientation of the gamma-tubulin small complex components using protein tags as labels for single particle EM.” *J Struct Biol* 168 (3): 571–4.

Hoopmann, M. R. et al. 2015. “Kojak: efficient analysis of chemically cross-linked protein complexes.” *J Proteome Res* 14 (5): 2190–8.

Kall, L. et al. 2007. “Semi-supervised learning for peptide identification from shotgun proteomics datasets.” *Nat Methods* 4 (11): 923–5.

Kollman, J. M. et al. 2011. “Microtubule nucleation by gamma-tubulin complexes.” *Nat Rev Mol Cell Biol* 12 (11): 709–21.

Kollman, J. M. et al. 2015. “Ring closure activates yeast gammaTuRC for species-specific microtubule nucleation.” *Nat Struct Mol Biol* 22 (2): 132–7.

Kollman, J. M. et al. 2008. “The structure of the gamma-tubulin small complex: implications of its architecture and flexibility for microtubule nucleation.” *Mol Biol Cell* 19 (1): 207–15.

Kollman, J. M. et al. 2010. “Microtubule nucleating Γ-TuSC assembles structures with 13-fold microtubule-like symmetry.” *Nature* 466 (7308): 879–882.

Pei, J., Tang, M., and Grishin, N. V. 2008. “PROMALS3D web server for accurate multiple protein sequence and structure alignments.” *Nucleic Acids Res* 36 (Web Server issue): W30–4.

Pettersen, E. F. et al. 2004. “UCSF Chimera–a visualization system for exploratory research and analysis.” *J Comput Chem* 25 (13): 1605–12.

Rieping, W., Habeck, M., and Nilges, M. 2005. “Inferential structure determination.” *Science* 309 (5732): 303–6.

Russel, D. et al. 2012. “Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies.” *PLoS Biol* 10 (1): e1001244.

Sali, A. et al. 2015. “Outcome of the First wwPDB Hybrid/Integrative Methods Task Force Workshop.” *Structure* 23 (7): 1156–67.

Schneidman-Duhovny, D., Pellarin, R., and Sali, A. 2014. “Uncertainty in integrative structural modeling.” *Current opinion in structural biology* 28:96–104.

Topf, M. et al. 2008. “Protein structure fitting and refinement guided by cryo-EM density.” *Structure* 16 (2): 295–307.

Topf, M. et al. 2006. “Refinement of protein structures by iterative comparative modeling and CryoEM density fitting.” *J Mol Biol* 357 (5): 1655–68.

Trabuco, L. G. et al. 2008. “Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics.” *Structure* 16 (5): 673–83.

Ward, A. B., Sali, A., and Wilson, I. A. 2013. “Integrative structural biology.” *Science* 339 (6122): 913–915.

Zelter, A. et al. 2010. “Isotope signatures allow identification of chemically cross-linked peptides by mass spectrometry: a novel method to determine interresidue distances in protein structures through cross-linking.” *J Proteome Res* 9 (7): 3583–9.

Appendix 1:

EM Methods for the molecular architecture of the yeast mediator complex

Summary

The Megadalton Eukaryotic Mediator complex plays an essential role in transcription initiation acting as a scaffold for the recruitment of the general transcription apparatus and communication of regulatory signals via direct interaction with DNA bound activators. Currently, our structural understanding of the complex is limited: roughly one third of the complex has been determined at high resolution but the remainder is limited to coarse subunit mapping onto 2D EM projections. In this project, we combined an extensive crosslink dataset with 3D EM restraint in an integrative modeling approach to produce a detailed 3D subunit architecture map of the *S. cerevisiae* Mediator complex. Our modeling results provide a number of striking new architectural features such as the important roles of Med17 and Med14 in connecting Mediator Modules. Furthermore, we presented a detailed description of the domain-level localization and organization of Mediator subunits such as the colocalization of the N-termini of Med2 and Med3 in a probable coiled coil motif and the scaffolding role of a seven- bladed Beta propeller within the N-terminus of Med16. The model was highly consistent with both structural and proteomics data that were not employed as input modeling restraints.

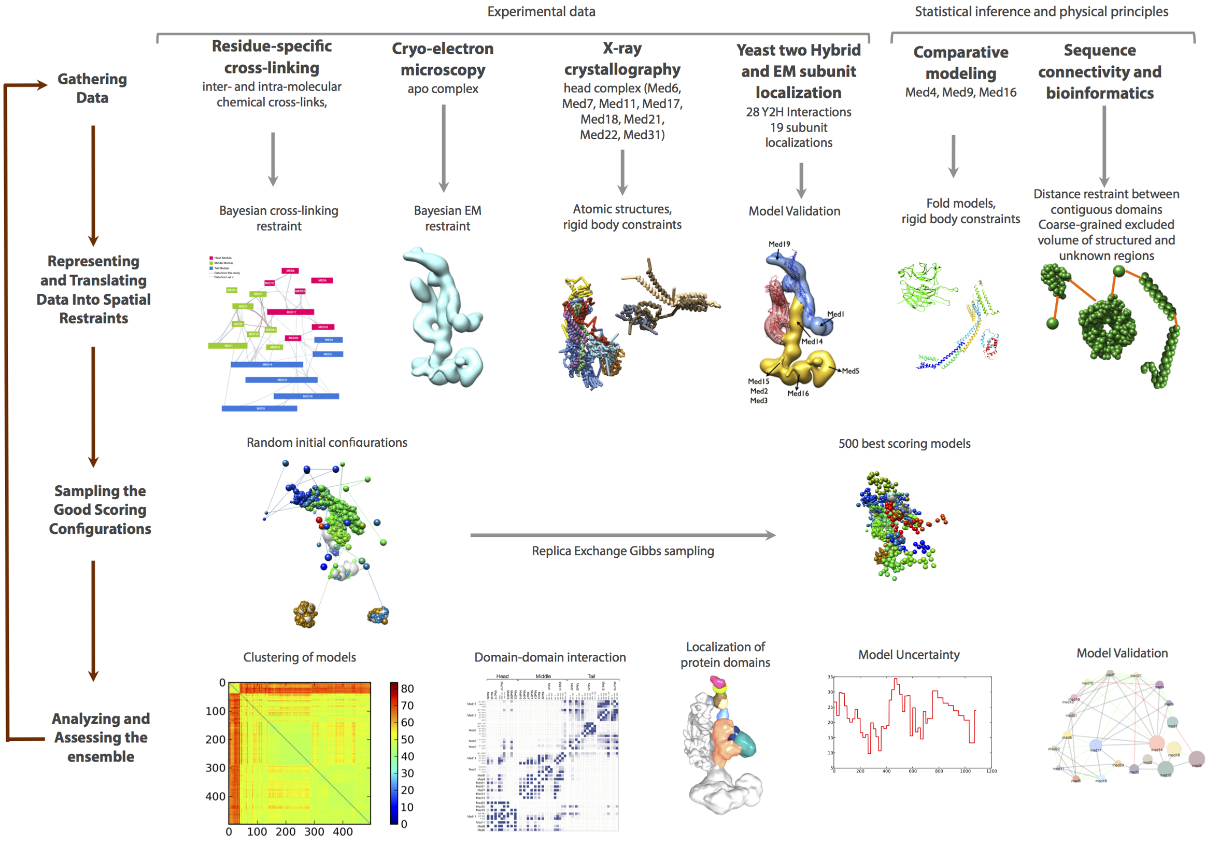
In this Appendix, we only report the EM methods relevant to the overall thesis. The approach was designed for fitting large systems into their EM density maps, particularly in the case of large rigid bodies. The idea is to approximate the electron density of both the model and the target with a Gaussian Mixture Model (GMM) and then use an analytic overlap function originally developed in Kawabata 2008. The advantage is that the score is quite fast to calculate, as no map simulation is required.

Representation

To maximize computational efficiency while avoiding model oversimplification, we represented the Mediator complex subunits in a multi-scale fashion (**Figure 1**). This representation employed spherical beads of varying sizes and 3D Gaussians, which coarsen domains of the complex using several resolution scales simultaneously. The spatial restraints discussed below are applied to individual resolution scales as appropriate (below). To improve the accuracy and precision of the model ensemble obtained by satisfaction of spatial restraints, beads and Gaussians of a given domain were arranged into either a rigid-body or a flexible string, based on the available crystallographic structures and comparative models. In a rigid-body, all the beads and the Gaussians of a given domain have their relative distances constrained during configurational sampling, while in a flexible string the beads and the Gaussians are restrained by the sequence connectivity.

Three scales were used to represent the crystallographic structures and the comparative models. Two scales, here named fine and coarse, were determined by applying two different coarse-graining criteria to the molecular volume of the atomistic structure. In both representations the volume was decomposed into beads. For the fine scale, each bead corresponded to individual residues, and was centered at the position of the Cα atom. For the coarse scale, each bead represented 10-residue segments and was positioned on the center of mass of all atoms of the corresponding segment. The third scale is the Gaussian mixture model (GMM) approximation of the atomic electron density of the corresponding structure. The atomistic structures were converted into a GMM by first sampling points from the density, and then fitting the sample using the EM algorithm implemented in scikit-learn (Buitinck et al. 2013). We set the number of Gaussians to the number of residues in a component divided by 50.

We adopted a two-scale representation for the flexible strings (i.e. domains without crystallographic structures or comparative models). Each element in the string corresponded to up to 40 residues and was represented by a bead (coarse scale) and a spherical Gaussian. The fine scale representation was omitted for computation efficiency. The radius of the bead and the variance of the Gaussians were suitably determined to describe the average molecular volume and the molecular electron density of polypeptide segments, respectively. The bead and the Gaussian centers were enforced to be identical.



Experimental data
Statistical inference and physical principles
Gathering Data
Residue-specific cross-linking
inter- and intra-molecular chemical cross-links,
Cryo-electron microscopy
apo complex
X-ray crystallography
head complex (Med6, Med7, Med11, Med17, Med18, Med21, Med22, Med31)
Yeast two Hybrid and EM subunit localization
28 Y2H Interactions
19 subunit localizations
Comparative modeling
Med4, Med9, Med16
Sequence connectivity and bioinformatics
Representing and Translating Data Into Spatial Restraints
Bayesian cross-linking restraint
Bayesian EM restraint
Atomic structures, rigid body constraints
Model Validation
Fold models, rigid body constraints
Distance restraint between contiguous domains
Coarse-grained excluded volume of structured and unknown regions
Med19
Med1
Med14
Med5
Med15
Med2
Med3
Med16
Sampling the Good Scoring Configurations
Random initial configurations
Replica Exchange Gibbs sampling
500 best scoring models
Analyzing and Assessing the ensemble
Clustering of models
0
100
200
300
400
0
100
200
300
400
80
70
60
50
40
30
20
10
0
Domain-domain interaction
Head
Middle
Tail
Localization of protein domains
Model Uncertainty
Model Validation

Figure 1: Schematic of the integrative modeling approach highlighting the individual data inputs and the four stages in our modeling strategy.

EM scoring

The EM 3D restraint was imposed on the Gaussian Mixture Model (GMM) representation of each domain, using the cross-correlation coefficient between GMM representations of the EM volume and model components (Kawabata 2008). The weights of each GMM component were normalized to the relative mass of the component to the mass of the module. The density of a molecule with a GMM is given by:

$$f(\mathbf{r}|\Theta) = \sum_{i=1}^{N} \pi_i \phi(r|\mu_i, \Sigma_i)$$

Here π_i are the mixing weights (normalized to 1) and $\phi(\mathbf{r}|\mu_i, \Sigma_i)$ is a Gaussian density function with mean μ_i and covariance Σ_i:

$$\phi(\mathbf{r}|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{3/2}|\Sigma_i|^{1/2}} \exp\left[-\frac{1}{2}(r-\mu_i)^{\intercal}\Sigma_i^{-1}(r-\mu_i)\right]$$

The GMM approximation for an electron density map (the data) can be calculated using the standard expectation-maximization approach using scikit-learn. The overlap function between the model (M) and the data (D) GMMs is defined by:

$$\begin{aligned} ov(\phi_M, \phi_D) &= \int \phi(\mathbf{r}|\mu_M, \Sigma_M)\phi(\mathbf{r}|\mu_D, \Sigma_D)dr \\ &= \frac{1}{(2\pi)^{3/2}|\Sigma_M + \Sigma_D|^{1/2}} \exp\left[-\frac{1}{2}(\mu_M - \mu_D)^{\intercal}(\Sigma_M + \Sigma_D)^{-1}(\mu_M - \mu_D)\right] \end{aligned}$$

That can be generalized as the overlap function between two GMMs:

$$ov(f_M, f_D) = \sum_{i=1}^{N_M}\sum_{j=1}^{N_D} \frac{1}{(2\pi)^{3/2}|\Sigma_{Mi} + \Sigma_{Dj}|^{1/2}} \exp\left[-\frac{1}{2}(\mu_{Mi} - \mu_{Dj})^{\intercal}(\Sigma_{Mi} + \Sigma_{Dj})^{-1}(\mu_{Mi} - \mu_{Dj})\right]$$

The cross-correlation function is, as suggested elsewhere (Sfikas et al. 2005):

$$\begin{aligned} CC(f_M, f_D) &= \frac{2\int f_M(x) f_D(x) dx}{\int f_M^2(x) + f_D^2(x) dx} \\ &= \frac{2ov(f_M, f_D)}{ov(f_M, f_M) + ov(f_D, f_D)} \end{aligned}$$

The negative logarithm of the cross-correlation is the EM score. We empirically found a scaling factor of ~ 100.

References

Buitinck, L. et al. 2013. “API design for machine learning software: experiences from the scikit-learn project.” *arXiv preprint arXiv:1309.0238*.

Kawabata, T. 2008. “Multiple subunit fitting into a low-resolution density map of a macromolecular complex using a gaussian mixture model.” *Biophys J* 95 (10): 4643–58.

Sfikas, G. et al. 2005. “An Analytic Distance Metric for Gaussian Mixture Models with Application in Image Retrieval.” In *Artificial Neural Networks: Formal Models and Their Applications – ICANN 2005*, 3697:835–840. Springer Berlin Heidelberg.

Appendix 2:

Ring closure activates yeast γTuRC for species-specific microtubule nucleation

Ring closure activates yeast γTuRC for species-specific microtubule nucleation

Justin M. Kollman[1,6], Charles H. Greenberg[2], Sam Li[1], Michelle Moritz[1], Alex Zelter[3], Kimberly K. Fong[3], Jose-Jesus Fernandez[4], Andrej Sali[2], John Kilmartin[5], Trisha N. Davis[3], David A. Agard[1]*

1. Department of Biochemistry and Biophysics and Howard Hughes Medical Institute, University of California at San Francisco, San Francisco, California 94158, USA

2. Department of Bioengineering and Therapeutic Sciences, Department of Pharmaceutical Chemistry, and California Institute for Quantitative Biosciences (QB3), University of California at San Francisco, San Francisco, California 94158, USA

3. Department of Biochemistry, University of Washington, Seattle, WA 98195, USA

4. Centro Nacional de Biotecnología, Consejo Superior de Investigaciones Cientificas (CSIC), 28049 Madrid, Spain

5. MRC Laboratory of Molecular Biology, Cambridge CB2 0QH, England

6. Present Address: Department of Biochemistry, University of Washington, Seattle, Wasthington 98195, USA

*Correspondence:
David A. Agard
University of California at San Francisco
Mission Bay, Genentech Hall
600 16th Street
Room S412D
San Francisco, CA 94158-2517
agard@msg.ucsf.edu

The γ-tubulin ring complex (γTuRC) is the primary microtubule nucleator in cells. γTuRC is assembled from repeating subunits of the γ-tubulin small complex (γTuSC) and is thought to function as a template by presenting a ring of γ-tubulins that mimic microtubule geometry. However, the yeast γTuRC structure showed each γTuSC is in an open conformation that prevents matching to microtubule symmetry. By contrast, here we show that γ-tubulin complexes are in a closed conformation when attached to microtubules. To confirm the functional importance of ring closure we trapped a closed state of the complex and determined its structure to show that the γ-tubulin ring precisely matches microtubule symmetry. Importantly, the closed state is a stronger nucleator, suggesting the switch between open and closed states provides a means for allosteric control of γTuRC activity. We also show that γTuRCs have a profound preference for tubulin from the same species. Lastly, the structure of the closed state allows detailed insight into the architecture of γTuSC and γTuRC.

Microtubule nucleation is mediated *in vivo* by γ-tubulin complexes, which allow cells to control both the location and timing of new microtubule growth. The conserved core of the nucleating machinery is the γ-tubulin small complex (γTuSC), a 300 kDa V-shaped structure with two copies of γ-tubulin and one each of the accessory proteins GCP2 and GCP3, which are distant homologs of each other. GCP2 and GCP3 form the elongated arms of the complex, with γ-tubulin at each tip of the V[1,2]. Low-resolution structural studies of isolated γTuSCs showed that it is flexible, with a hinge-like motion near the center of the GCP3 arm[2]. In most eukaryotes several other accessory proteins, GCP4-6, assemble with multiple γTuSCs to form the γ-tubulin ring complex (γTuRC)[3,4]. γTuRC has long been thought to function as a template, presenting a ring of γ-tubulins from which microtubules grow[3-7].

Saccharomyces cerevisiae lacks the γTuRC-specific accessory proteins found in other eukaryotes, and nucleates microtubules from γTuSC oligomers. These oligomers are anchored to the nuclear face of the spindle pole body by the coiled-coil protein Spc110[8,9]. Isolated γTuSCs have a weak propensity to self-assemble, and the N-terminal domain of Spc110 (Spc110$^{1\text{-}220}$) stabilizes these interactions to promote formation of extended spiral-shaped oligomers that have 13 γ-tubulins per turn[10], matching the protofilament number of *in vivo* microtubules[11]. Extended γTuSC polymers are not observed at spindle pole bodies[12], and although some estimates of the subunit number have been made[13], the overall size and organization of the functional γTuSC oligomer *in vivo* is unclear.

While the nucleation activity of yeast γTuSC is entirely dependent on its oligomerization, unexpectedly the oligomers are configured with microtubule-like γ-tubulin lateral contacts only between γTuSCs, while within each γTuSC the two γ-tubulins are held apart in an open conformation[2,10]. A consequence of the staggered lateral γ-tubulin interactions is a γTuSC ring with a pitch ~25 Å larger than that of microtubules (Fig. S1). This conformation seems inconsistent with γTuSC assemblies acting as efficient microtubule templates; indeed, MT nucleation experiments showed only a modest enhancement over background. Based on flexibility observed within individual γTuSCs[2], however, we proposed that an allosterically regulated conformational change could result in a precise match to microtubule geometry, forming a template with increased nucleating activity[5,10].

Here, we demonstrate that *in vivo* the minus ends of microtubules are anchored to the spindle pole body via a γTuSC ring that is in a closed conformation and has a defined number of subunits. While the *in vivo* drivers of ring closure are unknown, we could trap a closed state of γTuSC oligomers by disulfide crosslinking, and determined its structure at 6.9 Å resolution by cryo-EM. The closed state closely resembles the structure observed at minus ends, and conformational changes within each γTuSC result in a nearly perfect match between the γ-tubulin ring and thirteen protofilament microtubule geometries. The closed state is more active than the open state, confirming that γTuRC activity can be conformationally regulated. We also show that yeast γTuSC is much more active with yeast tubulin than with vertebrate tubulin, demonstrating the importance of species specificity in nucleating activity. The high-resolution structure of the closed state allowed us to generate a pseudo-atomic model that provides a more detailed view of the interactions of components within the γTuSC and the nature of assembly contacts between γTuSCs.

RESULTS

γTuSC binds microtubules in a closed conformation

A key question is whether or not γTuSCs can actually form a structure that better matches microtubule symmetry. To answer this, we determined the conformational state of γTuSC rings interacting with microtubules at the spindle pole. We examined the *in situ* structure of microtubule minus ends attached to the nuclear face of purified spindle pole bodies using cryo-electron tomography (Fig. 1a). We averaged 1156 individual capped microtubule minus ends to generate a structure at 38 Å resolution ($FSC_{0.5}$ criterion). The structure reveals an asymmetric cap in which individual γTuSC subunits can be discerned in most directions. The γTuSCs form a lock-washer shaped spiral that rises 120 Å (Fig. 1b), similar to the pitch of 13-protofilament microtubules[14]. This pitch is in contrast to the 147 Å pitch of reconstituted γTuSC-Spc110$^{1-220}$

spirals[10], indicating that when bound to microtubules *in vivo* the conformation of the γTuSC ring is different than that observed for our unbound rings *in vitro* (ref).

Spc110 promotes formation of a defined yeast γTuRC

The structure of capped minus ends also reveals the number of γTuSC oligomers *in vivo.* It has been assumed that functional γTuSC oligomers would have either six copies (with twelve γ-tubulins and a gap at the thirteenth position), seven copies (with an overlapping, inaccessible fourteenth γ-tubulin at the end), or be variable in number[6,15,16]. In the minus end structure the last γTuSC can be see to overlap vertically with the first γTuSC (Figure 1b), indicating that the ring is formed from seven subunits. Unlike these defined seven subunit rings, *in vitro* assembly of γTuSC and Spc110$^{1\text{-}220}$ yields extended spirals[10]. However, we found that a longer Spc110 fragment (residues 1-401), with an additional 180 residues of predicted coiled coil, limited γTuSC assemblies to single rings and smaller oligomers, suggesting Spc110 sterically interferes with addition of more than seven γTuSCs (Fig. 1c,d). Thus, Spc110 both promotes γTuSC assembly and limits oligomer size, forming well-defined yeast γTuRCs.

γTuSC oligomers trapped in a closed state

Since γTuSCs adopts a closed, microtubule-like geometry when interacting with microtubules, we sought to trap this closed state with engineered disulfide bonds to test the functional consequences of closure. Assuming that in a closed state γ-tubulin makes lateral contacts between the M- and H1-S2 loops similar to those observed in the microtubule lattice[14] or γ-tubulin crystals[17,18] (Fig. 2a), we designed four sets of paired cysteine mutations: Asn57/Gly288, Ser58/Gly288, Asp128/Ser291, and Arg161/Arg341. These sites were predicted to be juxtaposed at the inter-γTuSC assembly interface, but widely separated at the intra γTuSC interface (Fig. 2a). Thus, under oxidizing conditions a disulfide bond should readily form between γ-tubulins from adjacent γTuSCs, but would only form between γ-tubulins within the same γTuSC if that γTuSC sampled a closed state. Disulfide bond formation was only observed in the N57C/G288C and S58/G288C mutants. Of the two S58C/G288C was better behaved in terms of oligomeric assembly, so we pursued structural and functional characterization of this complex, which we refer to as γTuSC^{CC}. The γTuSC^{CC} mutations were introduced into yeast on the only copy of γ-tubulin, where they did not affect viability or metaphase spindle organization (Fig. S2). Thus, the mutations do not impair γTuSC function in the context of the cell.

Under reducing conditions individual γTuSC^{CC} had the same overall structure as the wildtype complex[2] (Fig. S3a). Wildtype γTuSC has a weak propensity to self-assemble into ring-like oligomers under a narrow range of salt concentrations (<200 mM KCl) and pH (6.4-7.0). After removal of reducing reagents, however, γTuSC^{CC} spontaneously assembled into large oligomers at pH 7.6 and 500 mM KCl, conditions that strongly disfavor spontaneous assembly of wildtype γTuSC (Fig. S3b). This suggested that inter-γTuSC disulfide bonds stabilize weak interactions between γTuSCs. SDS-PAGE of non-reduced γTuSC^{CC} revealed a ladder of cross-linked γ-tubulin oligomers, indicating that disulfide crosslinks had formed both within and between γTuSCs (Fig. S3c). The presence of γ-tubulin oligomers with greater than two chains means that both inter- and intra-γTuSC disulfide crosslinks were formed, indicating trapping of a novel conformation of γTuSC that allowed formation of intra-γTuSC disulfide.

Next, we co-purified the γTuSC^{CC}-Spc110$^{1-220}$ complex and observed that under reducing conditions it formed filaments similar to the wildtype complex. Under weakly oxidizing conditions, however, there were clearly two populations of filaments in the sample, one similar to the open wildtype filaments (Fig. 2b, red arrows), and a new, more compact or closed form with a different helical pitch (Fig. 2b, green arrows). The power spectrum of single filaments of the open form had a strong layer line at 147 Å, corresponding to the pitch of the wildtype filament, while the closed form had a layer line at 122 Å, very close to the pitch of a 13 protofilament microtubule. Single filaments appeared to be predominantly of one form or the other, suggesting a cooperative transition from the open to closed morphology.

The structure of closed γTuSC oligomers

We determined the structure of the closed form of γTuSC^{CC}-Spc110$^{1\text{-}220}$ filaments by cryo-EM and iterative helical real space reconstruction[19]. To minimize heterogeneity in filament morphology γTuSC^{CC}-Spc110$^{1\text{-}220}$ was extensively dialyzed against an oxidizing buffer to promote full disulfide crosslinking (Fig. S3d,e). An initial structure was determined at 9.4 Å resolution, from about 94,400 γTuSC subunits. To eliminate residual open/closed heterogeneity in the filaments we sorted the dataset by comparing filament segments to the initial γTuSC^{CC} and the open wildtype structures, and omitted segments that matched better to the open structure (Fig. S3f). This yielded a final dataset with about 76,000 γTuSC subunits, and a structure at 6.9 Å resolution (Fig. S3g); the map appears to have anisotropic resolution, with the core density (N-terminal domains of GCP2/3) more well defined than peripheral densities (C-terminal domains of GCP2/3 and γ-tubulin).

The refined helical symmetry parameters for the closed structure (-55.4° rotation and 18.8 Å rise per subunit, versus -54.3°/22.2 Å for the open structure) match 13-protofilament microtubule geometry (Fig. 3a), and are a close match to the helical pitch of the subtomogram average of capped minus ends from spindle pole bodies (Fig. 1b). Moreover, in contrast to the staggered pairs of γ-tubulins in the open state, the intra- and inter-γTuSC lateral contacts are very similar to each other, and closely resemble the microtubule lattice (Fig. 3b). Thus, in a single turn of the closed state, the ring of γ-tubulins is positioned to act as a nearly perfect microtubule template.

The closed conformation enhances nucleation activity

The similarity to the microtubule lattice suggested that the nucleating activity of γTuSC would be enhanced in the closed state. We tested the difference in activity between the open and closed states with solution nucleation assays. We previously tested the nucleation activity of γTuSC in the context of filaments formed in the presence of Spc110$^{1\text{-}220}$ [10]. This likely underestimated the full activity because only γTuSCs at the end of a filament would have been accessible for interaction with α/β-tubulin. Here, we used individual γTuRCs in which all of the γ-tubulin is available for interaction with α/β-tubulin, formed with the longer Spc110$^{1\text{-}401}$ construct that promotes assembly but blocks filament extension (Fig. 1d).

We compared nucleating activity of reduced and oxidized single rings of γTuSC-Spc110$^{1\text{-}401}$ and γTuSC^{CC}-Spc110$^{1\text{-}401}$. Assays performed with mammalian brain tubulin showed only a 2-3 fold increase in activity over buffer controls when using assembled yeast γTuSCs[10]. Given the quality of the symmetry match, we became concerned that there could be some problem in using mammalian tubulin with yeast γTuSCs. So despite general assumptions of strong conservation of tubulins across species, we repeated these assays with wild-type γTuSC and purified recombinant *S. cerevisiae* α/β-tubulin[20]. With yeast tubulin we observed an approximately 300-fold increase in the number of microtubules (Fig. 3c,d), indicating a remarkably strong species-specificity in the γTuSC:α/β-tubulin interaction (Fig. S4a).

The activities of γTuSC and γTuSC^{CC} were tested under reducing (open) and oxidizing (closed) conditions. Wild-type γTuSC had similar nucleation rates under both conditions, while γTuSC^{CC} nucleation was similar to wildtype under reducing conditions but doubled under oxidizing conditions (Fig. 3c,d). Thus, artificially forcing the geometry of the γTuSC ring to better match that of the microtubule results in enhanced nucleating activity. While the improvement demonstrates the importance of geometric fidelity, the modest increase suggests that other conformational changes, such as a curved to straight transition within the γ-tubulin[19], may be required to fully activate the complex.

Pseudo-atomic model of γTuSC

We generated a pseudo-atomic model of γTuSC by fitting and refining homology models of γ-tubulin, GCP2, and GCP3 into the cryo-EM density of a single closed state γTuSC subunit (Fig. 4a,b) and the lower resolution open state we determined previously[10]. We faced several challenges in generating the model: the very low sequence identity between GCP2/3 with GCP4 (the only homolog with a crystal structure determined) made generating reliable homology models difficult, GCP2 and GCP3 are in different conformations than GCP4, and the anisotropic resolution of the EM structure led to less reliable fitting of the C-terminal domains of GCP2/3 and γ-tubulin. We developed several new tools to perform the modeling and to validate the model; the approach is outlined below in Methods, and a full description of the modeling procedure will be described elsewhere.

GCP2 and GCP3, which have similar overall shapes, were distinguished by previous labeling experiments[1]. The primary difference between the open and closed states is the degree of flexing in GCP2 and GCP3. Both change conformation, with GCP3 straightening by about 9° and GCP2 bending by about 8° in the closed state (Fig. S5a). In the closed state GCP2 and GCP3 are similar to each other and to the human GCP4 crystal structure, while GCP2 and GCP3 are in different flexed conformations in the open state (Fig. S5b).

Density in the EM map that we previously suggested corresponds to part of Spc110[10] is more easily resolved in the new, higher-resolution map and now clearly resembles a two-stranded coiled coil. The model includes 44 residues of Spc110 as a generic two-stranded coiled coil (Fig. 4, Fig. S6a). This density makes contacts with N-terminal regions of GCP2 and GCP3, with closest contacts near the base of GCP2 and the side of the coiled coil. Near the top (as depicted in Fig. S6), the Spc110 density splays apart, with one strand bending back toward GCP3, and the other extending away from GCP2 toward a γ-tubulin in the next layer of the helix. This model accounts for about 20% of the mass of the Spc110$^{1\text{-}220}$ construct. Much of the remainder of the molecule is likely near the center of the filament, where the ordered density on the outer surface connects to disordered density than runs through the core of the structure (Fig. S6b). This positioning of Spc110 is consistent with the fact that longer lengths of coiled coil inhibit filament elongation (Fig. 1d), as adding additional mass to the center of the filament would likely disrupt the helical packing.

The model also provides new insights into the contacts between γ-tubulin and GCP2/GCP3 (Fig 4b). Both GCP2 and GCP3 C-terminal domains make contacts with the minus end longitudinal surface of γ-tubulin, with the strongest contacts made with the T7 loop of γ-tubulin. The major difference between GCP2 and GCP3 in their interactions with γ-tubulin is an interaction between the H1-S2 loop of γ-tubulin with a loop in GCP3; the corresponding region of GCP2 is shifted away from γ-tubulin (Fig. 4b). The H1-S2 loop is at the lateral γ-tubulin interaction surface, and the interactions with GCP3 may serve to modulate lateral interactions.

DISCUSSION

Here, we have demonstrated that γTuSC can adopt a closed conformation upon which γ-tubulin becomes organized as a precise microtubule template (Fig. 3a,b). Flexing of both GCP2 and GCP3 in the closed state brings together the two γ-tubulins in each γTuSC with microtubule-like lateral contacts. GCP2 and GCP3 were modeled on the distant homolog GCP4. In the open state the conformations of both GCP2 and GCP3 differ from GCP4, whereas in the closed state all three are in similar conformations. This suggests that specific evolutionary changes in GCP2 and GCP3 have stabilized their open, less active, conformations. The advantage of stabilizing the lower-activity state would be the ability to rapidly and specifically upregulate γTuRC activity through allosteric activation. What remains to be determined is what factor or factors might induce conversion from the open to closed state.

Consistent with the idea that γTuSC conformation plays a role in modulating nucleating activity, the closed state is about twice as active as the open state (Fig. 3c). Importantly, yeast γTuRC demonstrated strong species specificity, being about 100-fold more active with yeast tubulin than with vertebrate tubulin. Previous assays of γTuSC and γTuRC activity, all performed with mammalian brain tubulin, found animal γTuRC to be much more active than the yeast counterpart[3,10,21,22]. However, our current results using yeast γTuRC with yeast tubulin show levels of activity similar to those of the animal complex with animal tubulin, emphasizing the importance of pairing tubulin and γTuRC from the same species in microtubule nucleation assays (Fig. S4a). A number of amino acid substitutions between vertebrate and yeast sequences at the α-tubulin to γ-tubulin contact surfaces may account for the species specificity (Fig. S4b,c).

The modest two-fold enhancement of activity is in line with the enhancement levels previously observed with γTuRC activating proteins in metazoans, where *in vitro* assays using isolated human γTuRCs in the presence of CDK5RAP2 activating proteins resulted in 5-fold enhancement of activity. Overexpression or depletion of CDK5RAP2 resulted in large-scale changes to the microtubule cytoskeleton in cells, suggesting that small shifts in the activity of the nucleating complex can have a large impact on cytoskeletal organization[22]. It is also possible that, in this artificial cross-linked system, other potentially important allosteric activations are not being recapitulated. For example, α/β-tubulin undergoes a dramatic curved-to-straight conformational change on polymerization[23,24], and a similar transition may accompany normal closure of γTuSC but not be induced in our cross-linked system, thus underestimating the enhancement of nucleation in a natively closed structure.

The γTuSC pseudo-atomic model was expanded by the helical symmetry of the closed γTuSC^{CC} filament to generate a model of the full γTuRC (Fig. 5a). We extrapolated from this to create a structural model of γTuRC bound to the minus end of a microtubule, assuming that longitudinal contacts between γ-tubulin and α-tubulin are similar to the α- to β-tubulin contacts (Fig. 5b). This model, in both overall morphology and helical symmetry, is remarkably similar to the microtubule minus end we observed in SPBs. In the model, thirteen γ-tubulins make contact with the microtubule; a fourteenth γ-tubulin, bound to GCP2 in the terminal γTuSC, lies directly below the first γ-tubulin, but does not make contact with the microtubule. Intriguingly, a novel interaction appears between one end of the ordered Spc110 density and the H6-H7 loop of the fourteenth γ-tubulin, suggesting a possible role for Spc110 in stabilizing the γTuRC ring at this overlapping position (Fig. 5c). In addition to the 13 longitudinal contacts between γ- and α-tubulin around the ring, the model predicts a single lateral contact between α-tubulin and γ-tubulin at the first position in the ring; a lateral interaction between the GCP3 C-terminal domain and an adjacent β-tubulin may also occur here (Fig 5d).

A large number of phosphorylation sites have been identified on γ-tubulin, GCP2 and GCP3 in *S. cerevisiae*[25-30]. Phosphorylation at a few sites has been shown to affect organization of the mitotic spindle[28,30,31], but the functional roles of most sites remain unknown. Mapping the sites on the γTuRC model shows they fall broadly into three categories: on the exterior of the γTuRC, at the exposed plus end of γ-tubulin, and on the interior of the γTuRC (Fig. S7). Exterior sites may be involved in modulating γTuRC interaction with binding partners, potentially affecting localization or activation. Sites at the plus end of γ-tubulin would be expected to directly modulate interactions with α/β-tubulin; similarly, sites at the lateral contact surfaces between GCP3 or γ-tubulin and α/β-tubulin may directly influence microtubule interactions (Fig. 5d). The phosphorylation sites on the ring interior, lying mostly near the minus end of γ-tubulin, would likely be inaccessible to kinases when γTuRC is bound to a microtubule, suggesting that phosphorylation at these sites occurs when γTuRCs are unoccupied, or possibly prior to assembly of γTuSCs into γTuRC and may serve to down regulate nucleation activity.

Here, we have demonstrated that γTuSC can adopt a closed conformation that closely resembles the morphology of capped microtubule ends *in vivo* and is more active for microtubule nucleation than the previously observed open conformation. The transition from an open to a closed state likely provides an allosteric mechanism for modulating γTuSC activity. It remains to be seen what factor or factors are involved in promoting this transition *in vivo*, and post-translational modification of γTuSC components[28,31] or direct binding by regulatory proteins may be required[22,32,33]. Regulation of γTuRC at the levels of γTuSC assembly, post-translational modification, and open-to-closed conformation are likely all involved in providing precise control of nucleating activity.

ACKNOWLEDGEMENTS

The structures of the capped microtubule end and γTuSC^{CC}-Spc110$^{1\text{-}220}$ filament have been deposited in the EM data bank (accession numbers EMD-5989 and EMD-2799, respectively). The authors are grateful to Michael Braunfeld and Xueming Li for assistance with cryo-EM data collection, and Luke Rice for help with over-expression and purification of yeast tubulin. This work was funded by NIH grant R01 GM040506 to TND and P01GM105537 to Mark Winey. Kim Fong was funded by T32 GM007270.

METHODS

Spindle pole body purification and imaging. Spindle pole bodies from *Saccharomyces cerevisiae* were purified following previous published procedures [34,35]. The purified SPB sample, initially in high concentration of sucrose, was first dialyzed at 4ºC overnight in a buffer containing 10 mM Bis-Tris/Cl (pH=6.5), 0.1 mM $MgCl_2$, 20% (v/v) DMSO. Next day, after mixing with 10 nm colloid gold, the sample was applied onto either a home-made holey carbon grid or a Quantifoil grid (PSI, Inc.) in a humidity chamber, then blotted and plunged into liquid ethane using a home-made plunger or a Vitrobot (FEI, Inc.). Frozen grids were stored in liquid nitrogen before use. Tomography data were collected on a Polara electron microscope (FEI, Inc.) running at 300kV. A post-column energy filter (GIF, Gatan, Inc.) was used and the slit width was set at 25 eV. Automatic data collection was carried out by UCSF Tomography software [36]. Single-axis tilt series were collected at a nominal magnification of 41,000. Images of dimension 2032x2032 were recorded on a CCD camera (UltraCam, Gatan, Inc.). The final pixel size on the images was 5.32 Å. The specimen was tilted from -60º to +60º in 1.5º step. The microscope defocus values were set in the range of 10 to 15 μm. The accumulated dose for each tilt series was ~60 $e^-/Å^2$.

Subvolume averaging of capped minus ends. For averaging γ-TuRC, the minus-end caps of microtubules attached to the nuclear face of the spindle pole body were identified manually. A total of 1156 subtomograms containing the MT minus-end were selected from 61 tomograms. They were boxed out and subjected to reference-free alignment by either a maximum-likelihood algorithm in Xmipp package [37] or a constrained cross correlation algorithm [38]. The alignment was carried out progressively with improvement in resolution. The structure converged after about 5 rounds of iterations

Generation of γTuSC^{CC} and Spc110$^{1\text{-}401}$ expression vectors. Spc110$^{1\text{-}401}$ was generated using Spc110 and GST DNA from the Spc110-GST pFastBac vector[21]. Primers were designed to amplify the Spc110$^{(1\text{-}401)}$ coding sequence while adding: 1) a BamHI site immediately upstream of the open reading frame; 2) a PstI site immediately downstream of the 401st Spc110 codon. Primers were designed to amplify the GST coding sequence while adding: 1) a PstI site followed by a TEV cleavage site immediately upstream of the GST open reading frame; 2) a HindIII site immediately downstream of the GST stop codon. The resulting PCR products were cloned into the Invitrogen Zero Blunt TOPO vector according to the manufacturer's instructions. BamHI and PstI were used to excise Spc110$^{(1\text{-}401)}$ and HindIII and PstI were used to excise TEV-GST from their respective TOPO vectors. The resulting fragments were ligated into pFastBac (Invitrogen) linearized using HindIII and BamHI. The Bac-to-Bac baculovirus expression system (Invitrogen) was used to produce protein from Sf9 cells according to the manufacturer's instructions using Sf-900 II SFM liquid media (Invitrogen) supplemented with 2.5% fetal bovine serum.

To generate γTuSC^{CC}, S58 and G288 of γ-tubulin were mutated to cysteines using the QuikChange Multi Site-Directed Mutagenesis Kit (Agilent Technologies) according to the manufacturer's instructions. The resulting mutant sequence was PCR amplified and cloned into the Invitrogen Zero Blunt TOPO vector according to the manufacturer's instructions. XhoI and HindIII were used to excise the mutant Tub4 sequence. The resulting fragment was ligated into pFastBac (Invitrogen) linearized with the same enzymes.

For viability testing, integrating vectors based on the pRS306[39] backbone were constructed. These contained either wild-type or mutant (S58C G288C) Tub4. 432 bp of upstream and 334 bp of downstream flanking genomic DNA sequence was included in the vector surrounding the Tub4 coding sequence.

Preparation of recombinant γTuSC and γTuSC complexes. γTuSC or γTuSC^{CC} was co-expressed with GST-tagged Spc110 constructs in Sf9 cells and purified as described[10]. Briefly, cell lysate was incubated with glutathione resin, washed in H100 (40 mM Hepes PH 7.6, 100 mM KCl, 1 mM EGTA, 1mM $MgCl_2$), and eluted from the resin by cleavage of the GST tag with TEV protease as the final purification step. For cryo-EM γTuSC^{CC}-Spc110$^{1-220}$ filaments were at 2 mg/ml total protein in H100 and 1 mM oxidized gluathione. Negative stain samples were prepared as described[40] in 0.75% uranyl formate, and cryo-EM samples were prepared on C-FLAT holey carbon grids[41] using a Vitrobot (FEI Co.).

Imaging and three-dimensional reconstruction of γTuSC and γTuSC complexes. Negative stain samples were imaged on a Tecnai Spirit G^2 Biotwin electron microscope (FEI, Inc.) operating at 120 kV, and images were recorded on an Ultrascan 4000 CCD detector (Gatan, Inc.). Cryo-EM images were recorded on a Tecnai TF20 electron microscope operating at 200 kV, and images were recorded on 8k x 8k TemCam-F816 camera (TVIPS, GmbH) with a pixel size of 0.94 Å/pixel. Images were acquired in a defocus range of 0.8-2 μm. Defocus was determined with CTFFIND[42], and each micrograph was corrected by application of a Wiener filter. Particles were boxed out in 485 Å segments, overlapping by 448 Å. After several initial rounds of unrestrained alignment search the particles were centred with respect to the helix axis by integer pixel shifts.

Iterative helical real space reconstruction was performed essentially as described by Egelman[43] and Sasche, et al.[44], using SPIDER[45], using a low-pass filtered cylinder was used as the initial reference volume. Initial helical symmetry parameters were taken from the open state γTuSC filament (-54.3° rotation, 22.2 Å rise per subunit), and refined at each iteration with the program hsearch_lorentz[43]. Initial centering of the particles was carried out on 4x binned images, with subsequent refinement on 2x binned images. An initial structure was determined at ~9.6 Å resolution. To reduce open/closed heterogeneity in the data set cross correlations were calculated for each helical segment to the initial reconstruction and the original open state structure[10], and particles matching better to the open state were omitted from further rounds of refinement. After sorting five rounds of unrestrained alignment search were carried out with the 2x binned images. To minimize effects of bending in the helix each segment was masked to 200 Å along the helical axis and 340 Å perpendicular to the axis with a cosine-edged mask. These masked particles were then subjected to five rounds of local refinement of the unbinned images. Resolution was assessed by the $FSC_{0.5}$ criterion. Volumes were viewed and segmented using Chimera[46].

Fluorescence imaging of yeast cells carrying the γTuSC^{CC} mutation. All yeast strains were derived from W303. Fluorescent tags were introduced by PCR as described (http://depts.washington.edu/yeastrc "Plasmids and Protocols"), and *TUB4(S58C/G288C*) was integrated at the *TUB4* locus into *glr1Δ* cells, which carry a deletion for the gene encoding glutathione reductase and have high levels of oxidized glutathione[47]. The *glr1Δ* strain was used to increase the favorability of forming disulfides in the *TUB4* mutant. Live cells were mounted for microscopy on a 1% agarose pad[48]. Images were acquired at a single focal plane, with 1x1 binning, using a U Plan Apo 100× objective lens (1.35 NA), an Olympus IX70 inverted microscope, and a CoolSnap HQ digital camera (Photometrics) managed by softWorX software (Applied Precision). Exposures were 0.4 s for both mCherry and GFP. The images were processed as previously described[49] using custom Matlab programs (available upon request) to identify and quantify mCherry and GFP fluorescence intensities.

Yeast Strains Used:

All yeast strains also have *ade2-1oc ade3Δ-100 can1-100 his3-11,15 leu2-3,112 trp1-1 ura3-1* except as shown.

Yeast Strains

KFY36-13C	*MATalpha LEU2::GFP-TUB1 lys2Δ::HIS3 SPC42-mCherry::hphMX*
KFY42-1C	*MAT**a** LEU2::GFP-TUB1 lys2Δ::HIS3 glr1Δ::TRP1 SPC42-mCherry::hphMX*
KFY91	*MAT**a** LEU2::GFP-TUB1 lys2Δ::HIS3 glr1Δ::TRP1 SPC42-mCherry::hphMX TUB4(S58C/G288C)*
KFY135-8B	*MAT**a** lys2Δ::HIS3 TUB4(S58C/G288C)::URA3::tub4Δ::kanMX glr1Δ::TRP1 SPC42-mCherry::hphMX NUF2-GFP::kanMX*
KFY135-47A	*MAT**a** glr1Δ::TRP1 SPC42-mCherry::hphMX NUF2-GFP::kanMX*
KFY138-5A	*MAT**a** NUF2-GFP::kanMX SPC42-mCherry::hphMX*

Microtubule nucleation assays

Yeast tubulin was over-expressed and purified as described[20]. Pure γTuSC-Spc110, γTuSC alone, or buffer control (40mM K-HEPES pH6.9, 100mM KCl, 1mM EGTA, 1mM MgCl$_2$, 20% glycerol, 100µ M GTP, 100µ M oxidized or reduced glutathione) and *S. cerevisiae* tubulin were diluted at the appropriate concentrations into microtubule assembly buffer (80mM K-PIPES pH 6.9, 125mM KCl, 20% glycerol, 1mM EGTA, 1mM MgCl$_2$, 1mM GTP, 100µ M oxidized or reduced glutathione) on ice. Reactions were incubated at 30°C for 20 min, fixed 3 min in 10 volumes of 1% glutaraldehyde in BRB80 (80mM K-PIPES pH 6.9,1mM EGTA, 1mM MgCl$_2$), and then diluted 10 times into BRB80 (final volume 1.5ml). 1ml of the resulting fixed reactions was layered onto 20% glycerol/BRB80 cushions and centrifuged for 45 min, 24,000 x g, onto 18mm round coverslips. Microtubules were visualized on the coverslips by immunofluorescence with FITC-mouse-anti- ✔-tubulin (Sigma F2168) and 5-10 fields of microtubules were counted for each experiment.

Homology modeling and flexible fitting.

Models of γ-TuSC were computed using a combination of comparative protein structure modeling and flexible fitting into the EM density map of the closed state. To create a template structure for γ-TuSC, we rigidly docked 2 copies of the crystal structure of human GCP4 into the density map using UCSF Chimera[46]. Independently, initial alignment of the sequences in the TUBGCP family was performed with Promals3D[50]. Next, we produced an initial homology model of the Spc97-Spc98 dimer based on the initial alignment and the GCP4 template, using MODELLER 9.13[51]. The alignment was then iteratively refined by hand to improve the fit of the model into the density map. Using the final alignment, 200 homology models of the dimer were produced. We completed each dimer into a model of γ-TuSC by rigidly docking two copies of γ-tubulin and a coiled-coil fragment of Spc110. An additional neighboring copy of the complete γ-TuSC structure was added to model the inter-γ-TuSC interface. Subsequently, each γ-TuSC dimer was flexibly fitted into the density map using MDFF[52], with additional restraints to preserve helical symmetry, secondary structure, and conformation of the γ-tubulin domains. The best scoring model, as defined by the highest cross-correlation coefficient between the model and the map, was subjected to additional local sampling to estimate model precision, using MODELLER to randomize loops and MDFF to re-optimize the structures. The above process was repeated for the open state, using the open-state EM density map and the final alignment from the modeling of the closed state.

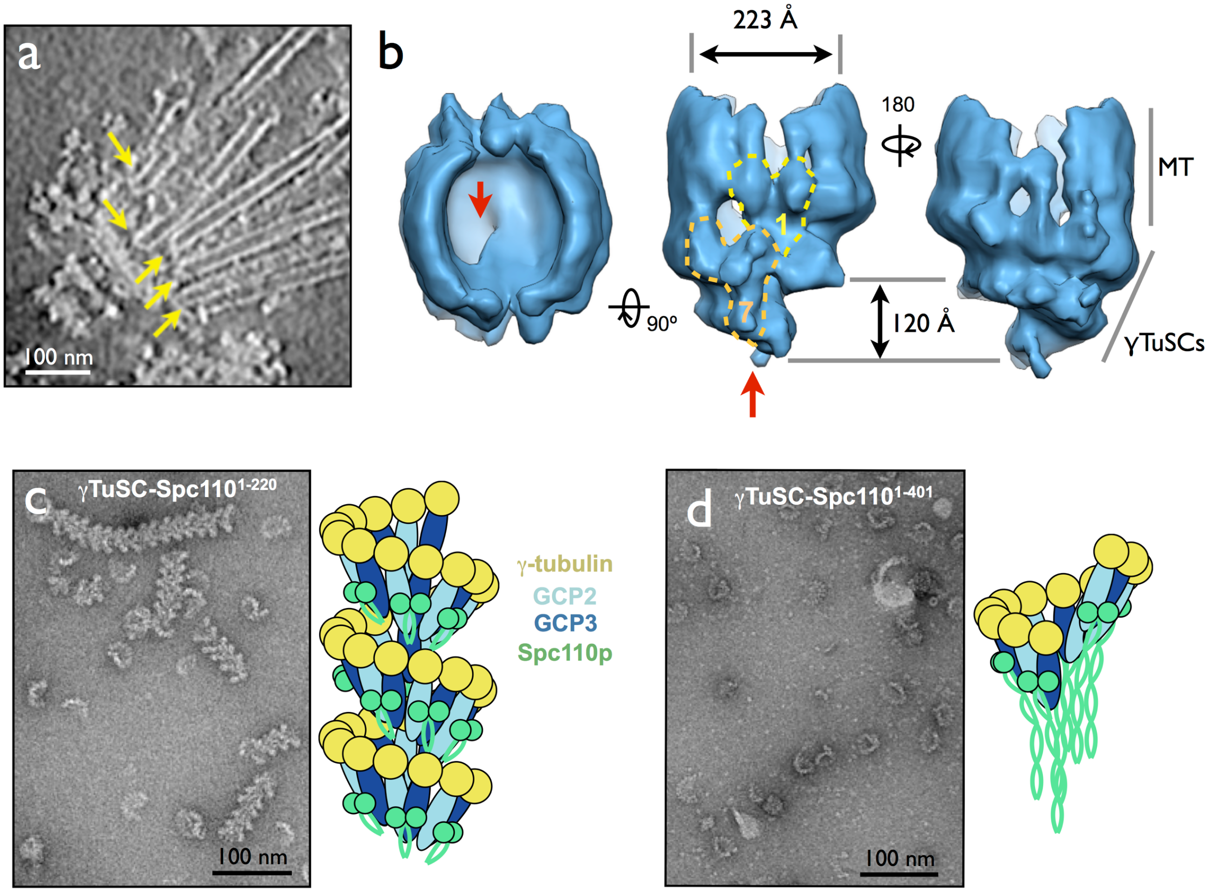


Figure 1. The yeast γTuRC is formed from seven γTuSCs and is limited in size by Spc110. a) A slice from a tomogram of isolated spindle pole bodies clearly shows the capped minus ends of microtubules. b) Subtomogram averaged structure of microtubule minus end. Red arrows indicate the position of the half-subunit overlap between the first and seventh γTuSC (outlined in yellow and orange, respectively). The 120 Å longitudinal rise of the γTuSC ring is indicated. c) γTuSC assembles extended filaments when bound to Spc110$^{1-220}$. Spc110 binds the outer surface of γTuSC, and fits within the groove of the filament (cartoon). d) Spc110$^{1-401}$ promotes assembly of γTuSC rings, but prevents extension beyond a single ring, suggesting that the longer predicted coiled-coil domain interferes with formation of oligomers greater than 7 γTuSC subunits.

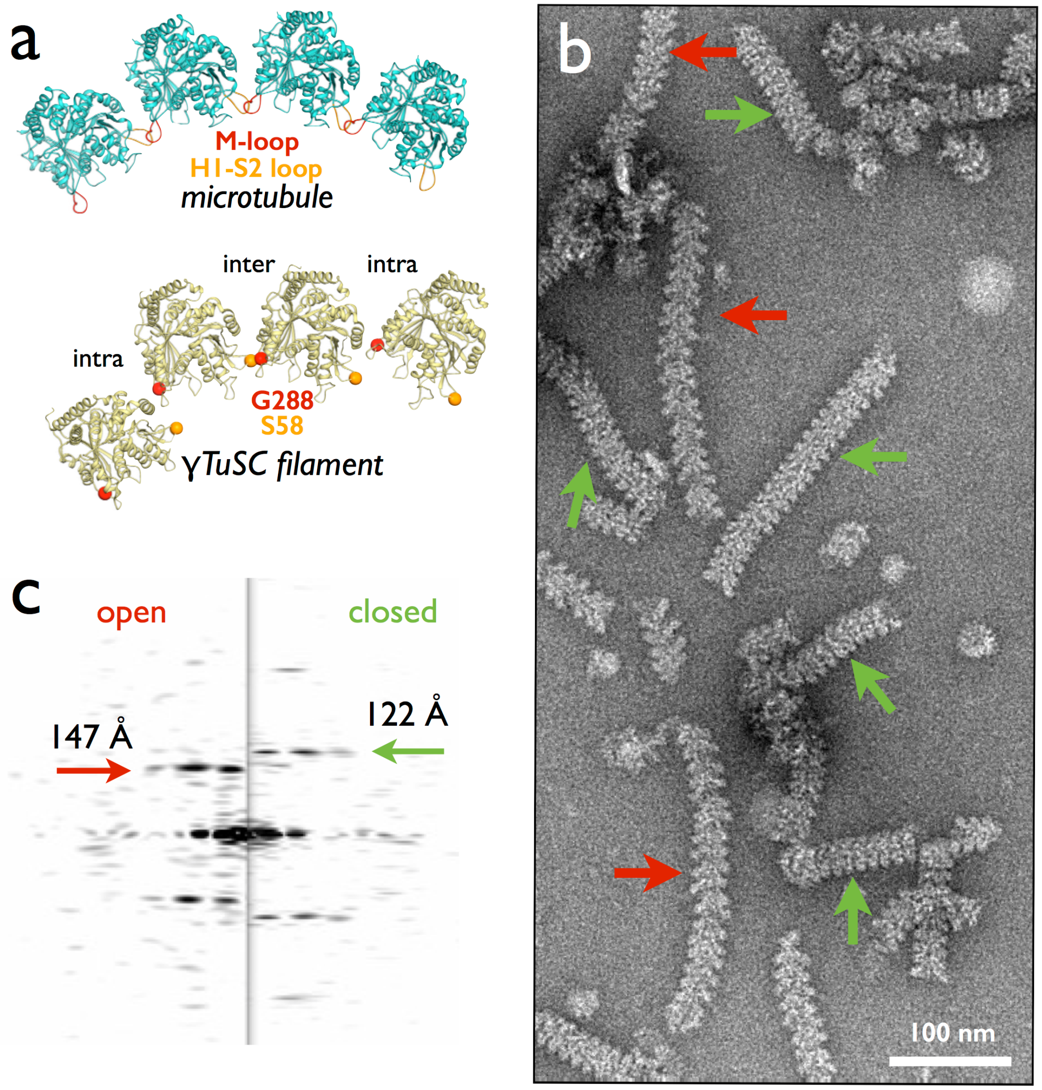


Figure 2. Engineered disulfides alter γTuSC filament morphology a) The lateral interface between β-tubulin subunits in the 13-protofilament microtubule, and the corresponding lateral interactions between γ-tubulins in the open state filament are shown. At the intra-γTuSC interface residues S58 and G288 are widely separated, but are close enough to potentially form a disulfide at the inter-γTuSC interface. b) Negative stain electron micrograph of the double mutant S58C G288C (γTuSC^{CC}) in complex with Spc110$^{1-220}$. The sample was dialyzed against an oxidizing buffer, after which two distinct filament morphologies were apparent (red and green arrows). c) Power spectra of individual filaments of different morphologies from (b) have different layer line spacing indicating different helical pitch. The more open morphology (red arrows in b) has a pitch of 147 Å, matching the previously reported structure of wildtype γTuSC, and the closed morphology (green arrows in b) has a pitch of 122 Å, matching the pitch of 13-protofilament microtubules.

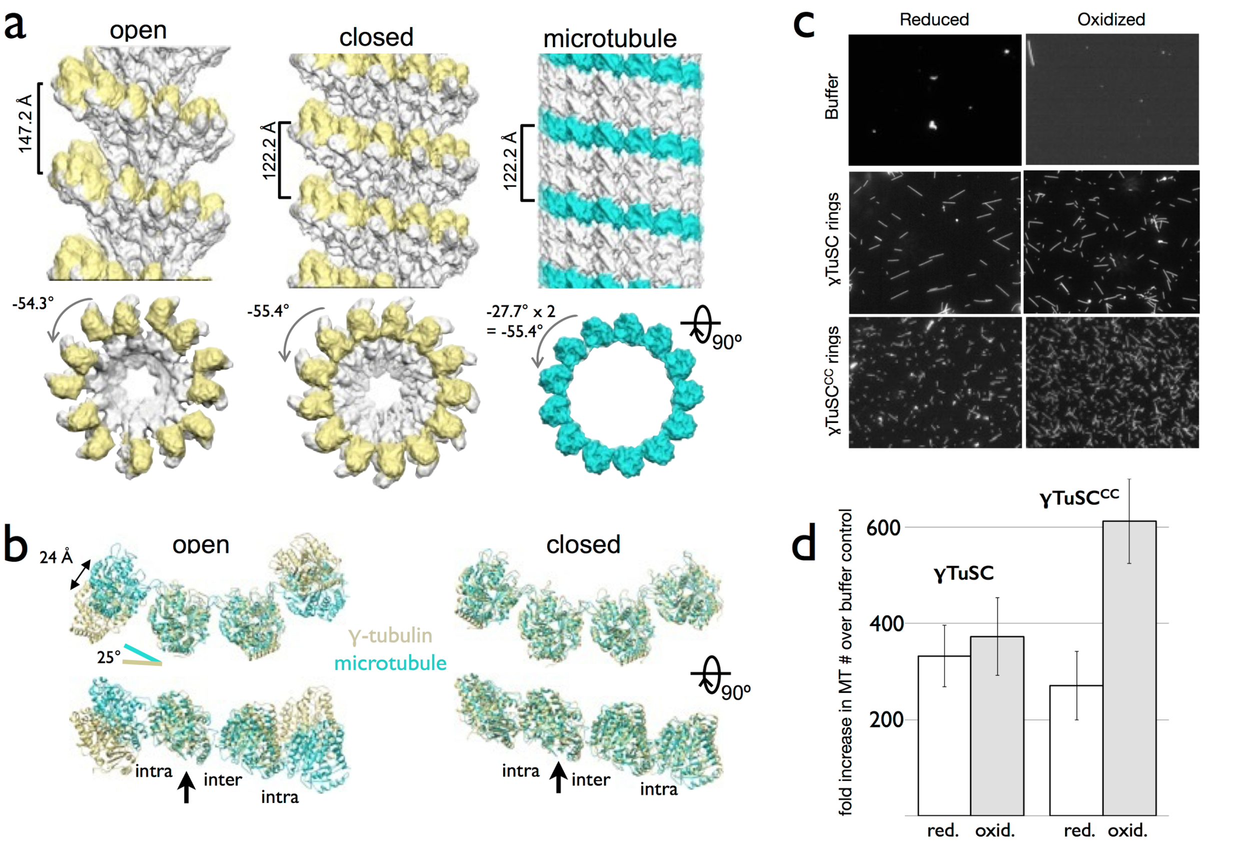


Figure 3. In the closed state γTuSC matches microtubule symmetry and has increased nucleation activity. a) The open state γTuSC filament, closed state γTuSC^{cc} filament, and 13-protofilament microtubule structure. γ-tubulin is highlighted in gold in the γTuSC structures, and the pitch of the 3-start helix in the microtubule is highlighted in cyan. Refined helical pitch and rotation per subunit are indicated. The rotation per subunit for the closed state γTuSC^{cc}, with two γ-tubulins, is exactly double the rotation between α/β tubulin subunits, indicating that the tubulin symmetries in the two structures match. b) The γ-tubulin crystal structure was fit into the γTuSC structures in the open and closed state. γ-tubulins from two adjacent γTuSCs are shown in gold, and the spacings between inter- and intra-γTuSC γ-tubulin pairs are indicated. The γ-tubulin indicated by the arrow was superimposed on a β-tubulin from the microtubule (cyan). In the open state (left) the intra-γTuSC γ-tubulin is displaced 25 Å from the corresponding position in the microtubule lattice. In the closed state (right) the γ-tubulin arrangement closely matches the β-tubulin lattice positions. c) Yeast tubulin (2.5 μM) was incubated at 30 °C with γTuSC-Spc110$^{1-401}$, γTuSC^{cc}-Spc110$^{1-401}$ complexes (100 nM) or buffer/γTuSC control under reducing or oxidizing buffer conditions. The resulting microtubules were fixed, centrifuged onto coverslips, visualized by immunofluorescence, and counted. d) Microtubules were counted for five fields per experiment (n=4), and the fold increase over buffer/γTuSC alone controls is plotted. Activity of γTuSCs alone was similar to buffer controls, with only a few microtubules on the entire coverslip (not shown). Error bars represent the standard error of the mean. T-tests confirm significant differences between activity of γTuSC110^{CC} complexes under oxidizing and reducing conditions (p ⦤ 0.013) , or between mutant and wild-type: (p ⦤ 0.055).

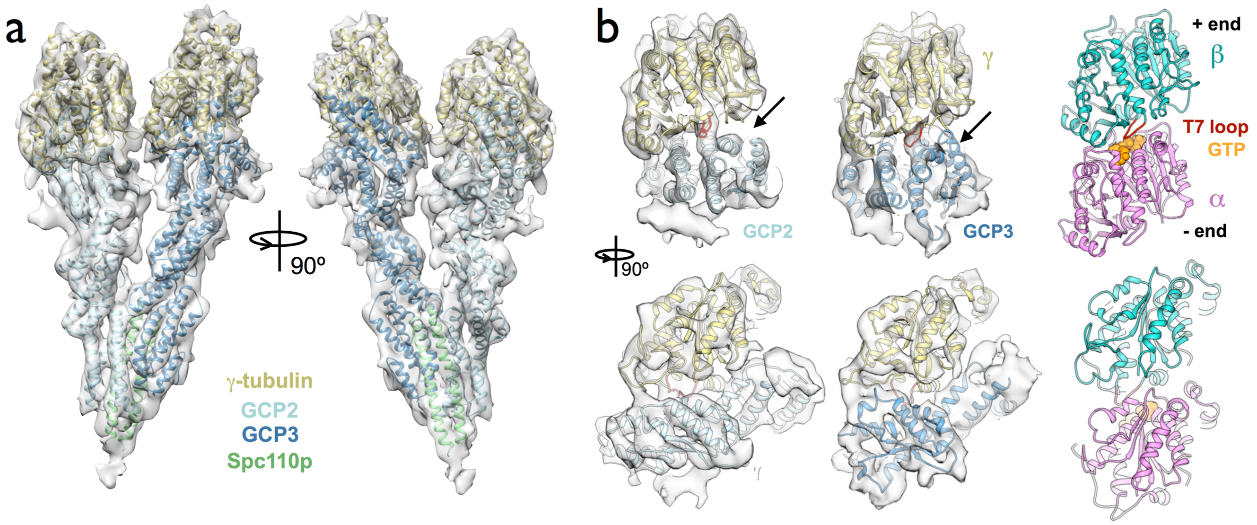


Figure 4. Pseudo-atomic model of γTuSC in the closed conformation. a) A pseduo-atomic model (ribbon diagram) was generated by molecular dynamics based flexible fitting of homology models of γTuSC components into the cryo-EM structure (semi-transparent surface). Spc110 was modeled as a generic length of coiled-coil. b) Close up views of the interactions of γ-tubulin with the C-terminal domains of GCP2 and GCP3, compared to longitudinal interactions within the α/β-tubulin heterodimer. The T7 loop (red) makes contact with GTP in α/β-tubulin and interacts with GCP2/GCP3 in γ-tubulin. The top view is the view from inside the microtubule, and the bottom view is looking at lateral interaction surfaces. In γTuSC both GCP2 and GCP3 bind the minus end of γ-tubulin, including the T7 loop. Contacts are made between the H1-S2 loop of γ-tubulin and residues 524-536 of GCP3; the corresponding region of GCP2 is shifted away from γ-tubulin in the GCP2 structure (arrows).

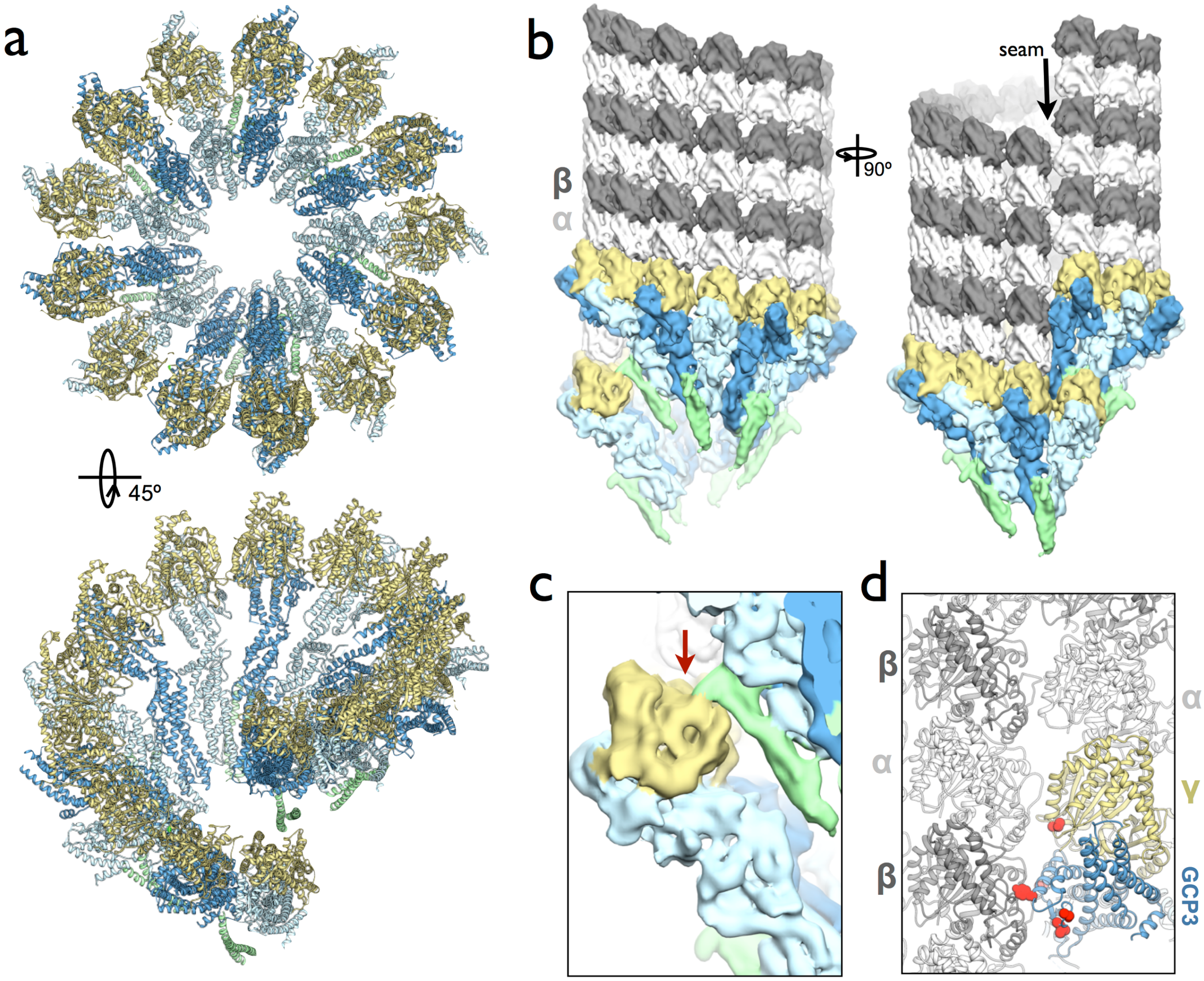


Figure 5. Pseudo-atomic model of γTuRC and its interactions with microtubules. a) The pseudo-atomic model of an complete yeast γTuRC with seven γTuSCs. b) A model of yeast γTuRC interacting with the minus end of a microtubule. c) A potential contact between the last γ-tubulin in the ring, which is not directly interacting with the microtubule, and Spc110 bound to the first γTuSC (arrow) is seen in the $\gamma TuSC^{CC}$ structure. d) Magnified view of interactions between the first γTuSC and the microtubule. A lateral contact is made between the first γ-tubulin in the ring and an adjacent α-tubulin. Just below this, a potential contact may be made between β-tubulin and the C-terminal domain of the first GCP3 in the γTuRC. Known phosphorylation sites on γTuSC that could potentially modulate these lateral interactions are indicated with red spheres.

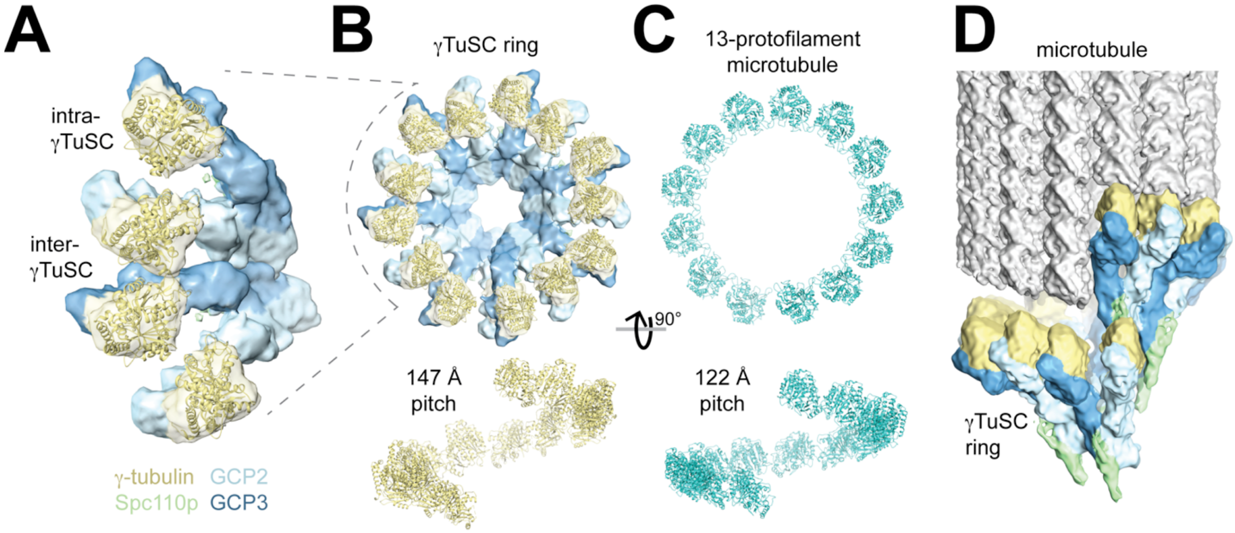
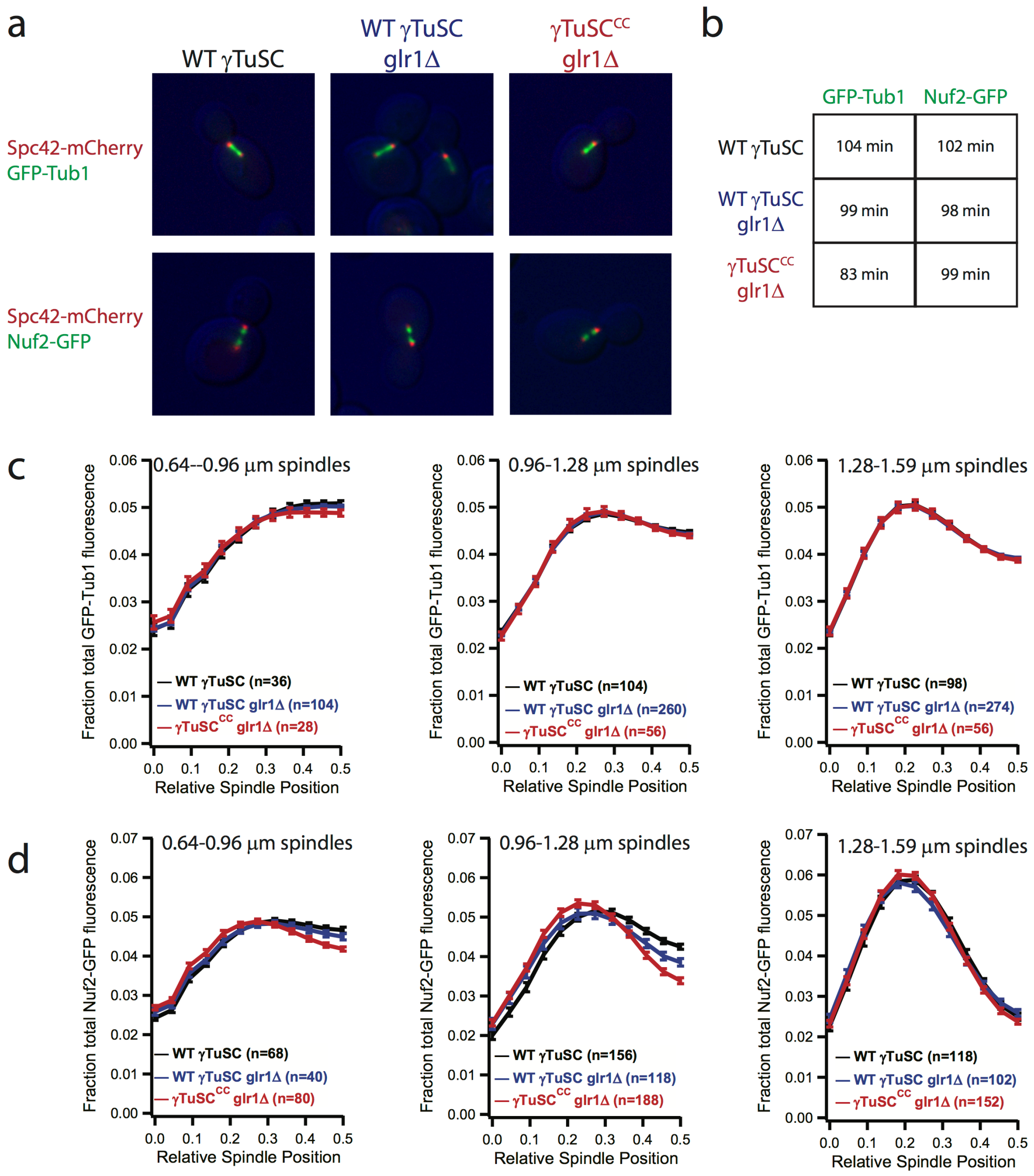


Figure S1. The structure of wildtype γTuSC oligomers does not perfectly match 13 protofilament microtubule geometry. a) Two laterally-associated γTuSC subunits from ring-like γTuSC assemblies are shown. The two γ-tubulins (gold) within each γTuSC are held separate from each other by GCP2 and GCP3, while the γ-tubulins at the inter-γTuSC interface are positioned with microtubule-like lateral contacts. b) The intra-γTuSC separation results in a staggered pattern of γ-tubulin in γTuSC assemblies, with a helical pitch of 147 Å. c) By contrast, lateral contacts between tubulin subunits in a microtubule are uniform around the filament, and have a helical pitch of 122 Å. d) To illustrate the mismatch between γTuSC oligomers and the microtubule, the upper γ-tubulin in the ring has been aligned to make longitudinal contacts with α-tubulin at the minus end of the microtubule. This would result in a large gap between α- and γ-tubulin at the last position in the ring.



	GFP-Tub1	Nuf2-GFP
WT γTuSC	104 min	102 min
WT γTuSC glr1Δ	99 min	98 min
γTuSC^{CC} glr1Δ	83 min	99 min

Figure S2. γTuSC^{CC} does not affect cell viability. a) Representative images of WT, glr1Δ, and γTuSCCC glr1Δ fluorescent strains. GLR1 encodes glutathione reductase; glr1Δ cells were used to maximize the chance that the γTuSC^{CC} disulfides would be oxidized. Cells in the top row have the spindle pole bodies marked by Spc42-mCherry and the microtubules marked by GFP-Tub1. Cells in the bottom row have the spindle pole bodies marked by Spc42-mCherry and the kinetochores marked by Nuf2-GFP. b) The doubling times of WT and γTuSC^{CC} strains. c) Average GFP-Tub1 fluorescence distributions show that γTuSCCC has no effect on tubulin distribution across three spindle length classifications. d) Average Nuf2-GFP fluorescence distributions show that γTuSCCC has no effect on kinetochore clustering in metaphase spindles (1.28-1.59 μm), while shorter spindles show kinetochores cluster slightly closer to the spindle pole bodies. Error bars represent the standard error of the mean.

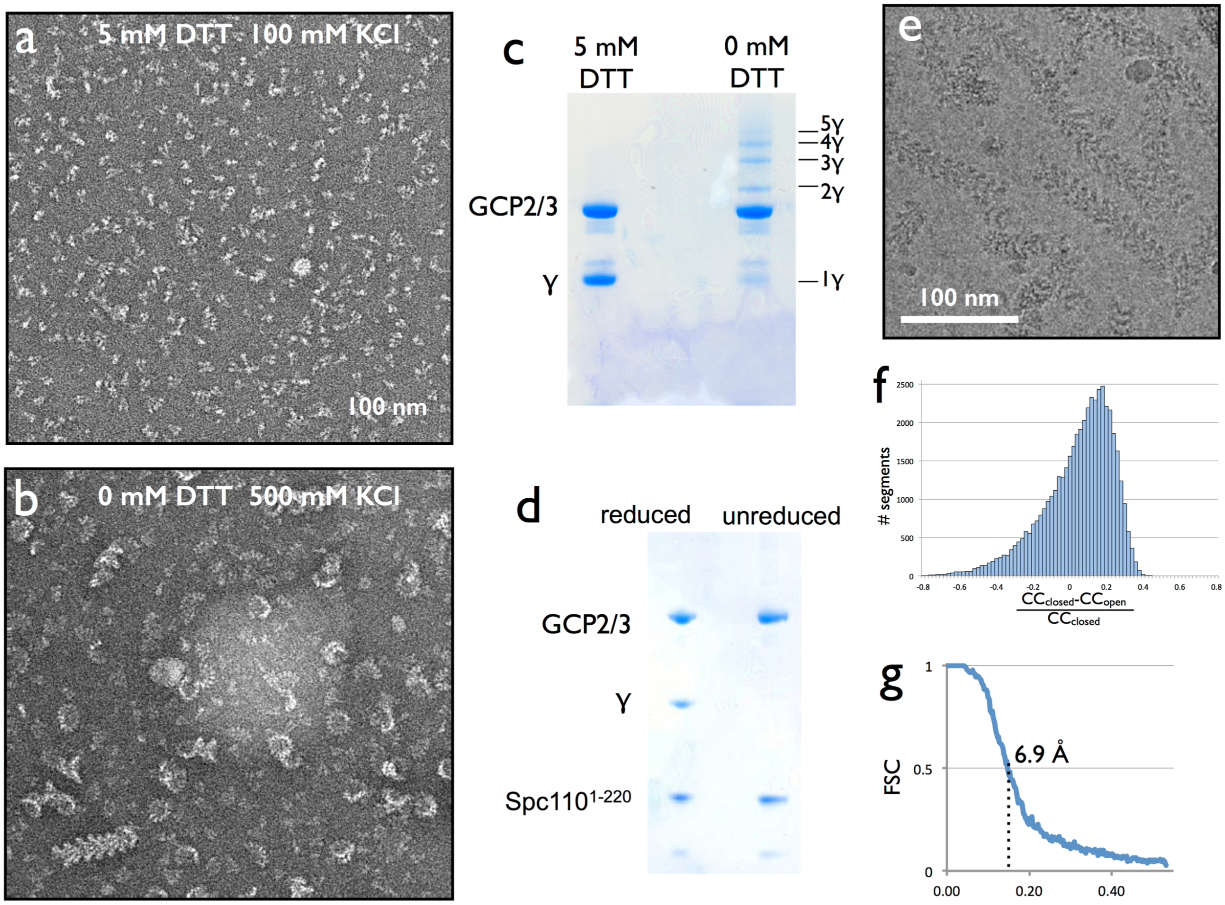
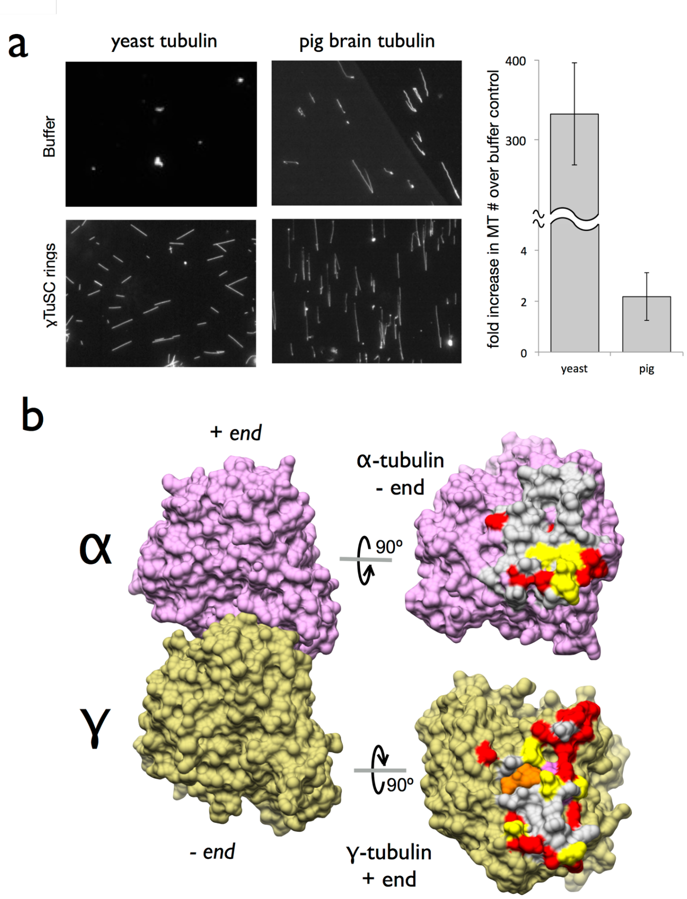


Figure S3. Crosslinking and structural analysis of γTuSC^{CC}. a) γTuSC^{CC} in the presence of 5 mM DTT is predominantly monomeric complex, although γTuSC pairs were frequently observed. b) Under non-reducing conditions γTuSCs spontaneously assemble into oligomers, even in the presence of 500 mM KCl, which prevents association of wildtype γTuSC. c) SDS-PAGE of γTuSC^{CC} under reducing conditions (5 mM DTT), and non-reducing conditions (0 mM DTT). In the non-reducing condition γ-tubulin forms disulfide crosslinks, resulting in a ladder of 1-5 γ-tubulin chains. Formation of γ-tubulin oligomers greater than two chains indicates that crosslinks were formed at both inter-γTuSC and intra-γTuSC interfaces. d) Co-purified γTuSC^{CC} and Spc110$^{1-220}$ was dialyzed for 72 h. against an oxidizing buffer containing 1mM oxidized glutathione. On SDS-PAGE the reduced sample has the expected bands for Spc110$^{1-220}$, γ-tubulin, and GCP2/GCP3, while in the unreduced sample extensive γ-tubulin crosslinking prevents its migration into the gel. This highly-cross-linked sample was used for cryo-EM imaging. e) Cryo-EM image of cross-linked γTuSC^{CC}-Spc110$^{1-220}$ filaments (same sample as in d). f) Segments of γTuSC^{CC} filaments were compared to projections of the γTuSC filament in the open state and a preliminary reconstruction of the closed state. The histogram plots the fractional difference in correlation coefficient, with negative values representing a better match to the open state. Segments matching better to the open state were omitted from the final refinement of the closed filament structure. c) Fourier shell correlation (FSC) of the final γTuSC^{CC}-Spc110$^{1-220}$ reconstruction indicates a resolution of 6.9 Å at the 0.5 cutoff.



```
yeast α    MREVISINVGQAGCQIGNACWELYSLEHGIKPDGHLEDGLSKPKGGEEGF 50
pig α      MRECISIHVGQAGVQIGNACWELYCLEHGIQPDGQMPSDK-TIGGGDDSF 49

yeast α    STFFHETGYGKFVPRAIYVDLEPNVIDEVRNGPYKDLFHPEQLISGKEDA 100
pig α      NTFFSETGAGKHVPRAVFVDLEPTVIDEVRTGTYRQLFHPEQLITGKEDA 99

yeast α    ANNYARGHYTVGREILGDVLDRIRKLADQCDGLQGFLFTHSLGGGTGSGL 150
pig α      ANNYARGHYTIGKEIIDLVLDRIRKLADQCTGLQGFSVFHSFGGGTGSGF 149

yeast α    GSLLLEELSAEYGKKSKLEFAVYPAPQVSTSVVEPYNTVLTTHTTLEHAD 200
pig α      TSLLMERLSVDYGKKSKLEFSIYPAPQVSTAVVEPYNSILTTHTTLEHSD 199

yeast α    CTFMVDNEAIYDMCKRNLDIPRPSFANLNNLIAQVVSSVTASLRFDGSLN 250
pig α      CAFMVDNEAIYDICRRNLDIERPTYTNLNRLIGQIVSSITASLRFDGALN 249

yeast α    VDLNEFQTNLVPYPRIHFPLVSYSPVLSKS----ESNSVSEITNACFEPG 296
pig α      VDLTEFQTNLVPYPRGHFPLATYAPVISAEKAYHEQLSVAEITNACFEPA 299

yeast α    NQMVKCDPRDGKYMATCLLYRGDVVTRDVQRAVEQVKNKKTVQLVDWCPT 346
pig α      NQMVKCDPRHGKYMACCLLYRGDVVPKDVNAAIATIKTKRTIQFVDWCPT 349

yeast α    GFKIGICYEPPTATPNSQLATVDRAVCMLSNTTSIAEAWKRIDRKFDLMY 396
pig α      GFKVGINYEPPTVVPGGDLAKVQRAVCMLSNTTAIAEAWARLDHKFDLMY 399

yeast α    AKRAFVHWYVGEGMEEGEFTEAREDLAALERDYIEVGAD- 435
pig α      AKRAFVHWYVGEGMEEGEFSEAREDMAALEKDYEEVGVDS 439


yeast γ     MGGEIITLQAGQCGNHVGKFLWSQLAKEHAIGTDGLSQLPDSSTERDDDT 50
human γ    --REIITLQLGQCGNQIGFEFWKQLCAEHGISPEAIVE--EFATEGTDRK 46

yeast γ    KPFFRENSRNKFTPRAIMMDSEPSVIADVE-NTFRGFFDPRNTWVASDGA 99
human γ    DVFFYQADDEHYIPRAVLLDLEPRVIHSILNSPYAKLYNPENIYLSEHG- 95

yeast γ    SAGNSWANGYDIGTRNQDDILNKIDKEIDSTDNFEGFQLLHSVAGGTGSG 149
human γ    -AGNNWASGFSQGEKIHEDIFDIIDREADGSDSLEGFVLCHSIAGGTGSG 144

yeast γ    LGSNLLEALCDRYPKKILTTYSVFPARS--SEVVVQSYNTILALRRLIED 197
human γ    LGSYLLERLNDRYPKKLVQTYSVFPNQDEMSDVVVQPYNSLLTLKRLTQN 194

yeast γ    SDATVVFDNASLLNISGKVFRNPNIDLQHTNQLISTIISSVTNSIRFPSY 247
human γ    ADCLVVLDNTALNRIATDRLHIQNPSFSQINQLVSTIMSASTTTLRYPGY 244

yeast γ    MYSSMSSIYSTLIPSPELHFLSPSFTPFTHS-SYDVMLDLLDPSNSLVST 296
human γ    MNNDLIGLIASLIPTPRLHFLMTGYTPLTKTTVLDVMRRLLQPKNVMVST 294

yeast γ    AMNNPTYFNVYNTIIGNVEPRQISRAMTKLQQRI-KFPSWSSSAMHVNIG 345
human γ    --TNHCYIAILNIIQGEVDPTQVHKSLQRIRERLANFIPWGPASIQVALS 342

yeast γ    RR-----SVSGMMLSNMSTVVNVFENACNTFDKVFAKGAFLNNYNVGDLF 390
human γ    RKSPYLPRVSGLMMANHTSISSLFERTCRQYDKLRKREAFLEQFRKEDMF 392

yeast γ    QSMQNVQDEFAESREVVQSLMEDYVAAEQDSYL-- 423
human γ    KDN---FDEMDTSREIVQQLIDEYHAATRPDYISW 424
```

Figure S4. Yeast γTuSC nucleation activity is species-specific. a) Comparison of nucleating activity of γTuSC rings with yeast and pig brain tubulin. Nucleation assays were performed as in Fig. 3C (data for yeast tubulin is the same as Fig. 3C), except that the experiments with brain tubulin were at 12 µM tubulin and 37° C. b) The predicted longitudinal interaction between α- and γ-tubulin is shown at left. Human γ-tubulin and pig α-tubulin are shown as these are the existing crystal structures. At right the interaction surfaces are highlighted as a function of the conservation between yeast and vertebrate tubulin: grey sites are identical, yellow have conservative amino acid substitutions, red have non-conservative substitutions, magenta is a two amino acid insertion in γ-tubulin, and GTP is shown in orange. c) Binary sequence alignments of yeast and vertebrate homologs, color coded as in (b).

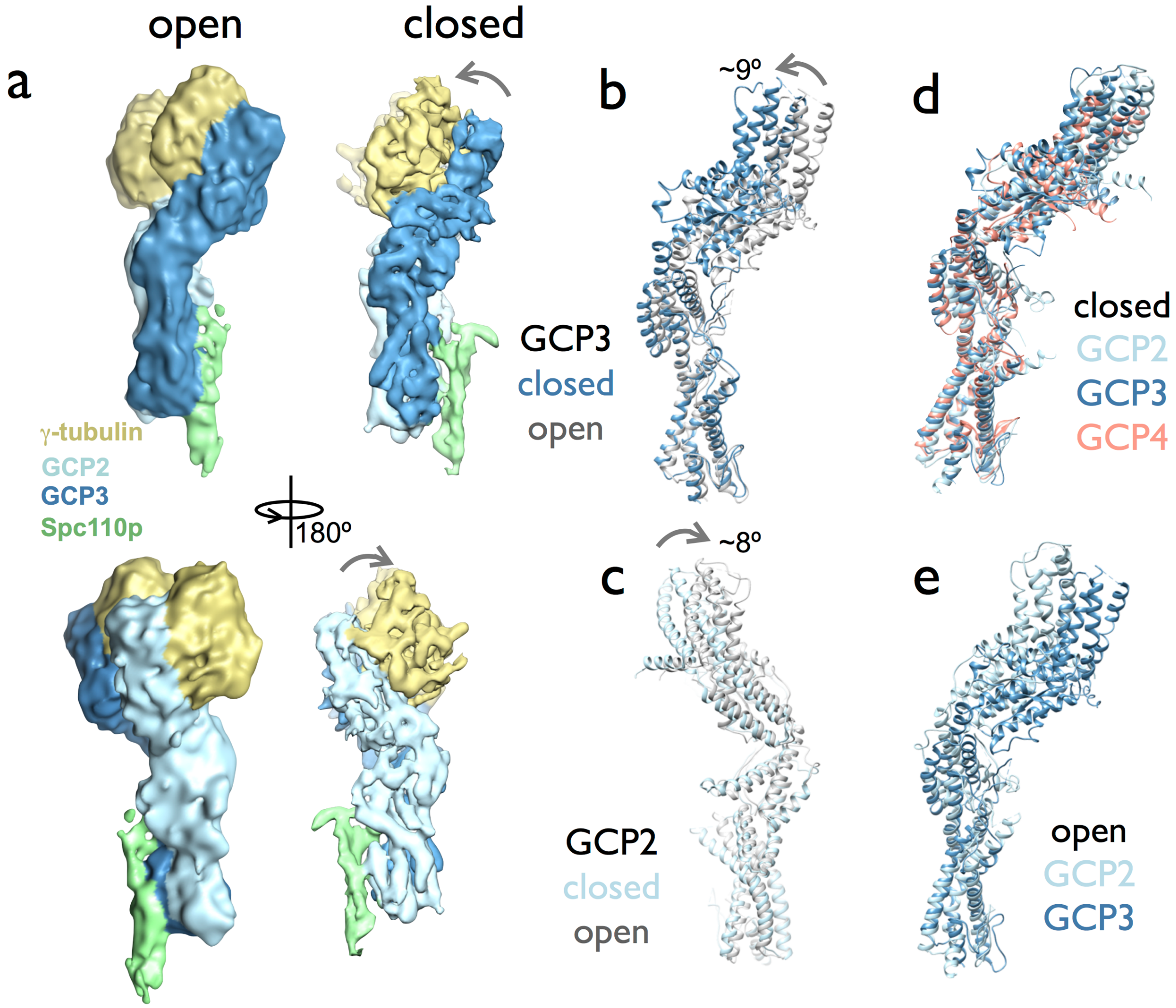


Figure S5. Differences in GCP2 and GCP3 in the open and closed states. a) Comparison of the open and closed states of γTuSC. The elongated GCP2 and GCP3 structures each appear slightly straightened. b) Overlay of GCP3 in the open (grey) and closed (blue) states shows that it straightens by about 9°. c) Overlay of GCP2 in the open (grey) and closed (blue) states shows that it straightens by about 8°. d) Superposition of GCP2 and GCP3 in the closed state shows that they are in very similar overall conformation. GCP4, which was used as the starting structure for homology modeling both GCP2 and GCP3, is in a similar overall conformation. e) Superposition of GCP2 and GCP3 in the open state. For B-E all structural alignments were done using the two N-terminal helical bundles.

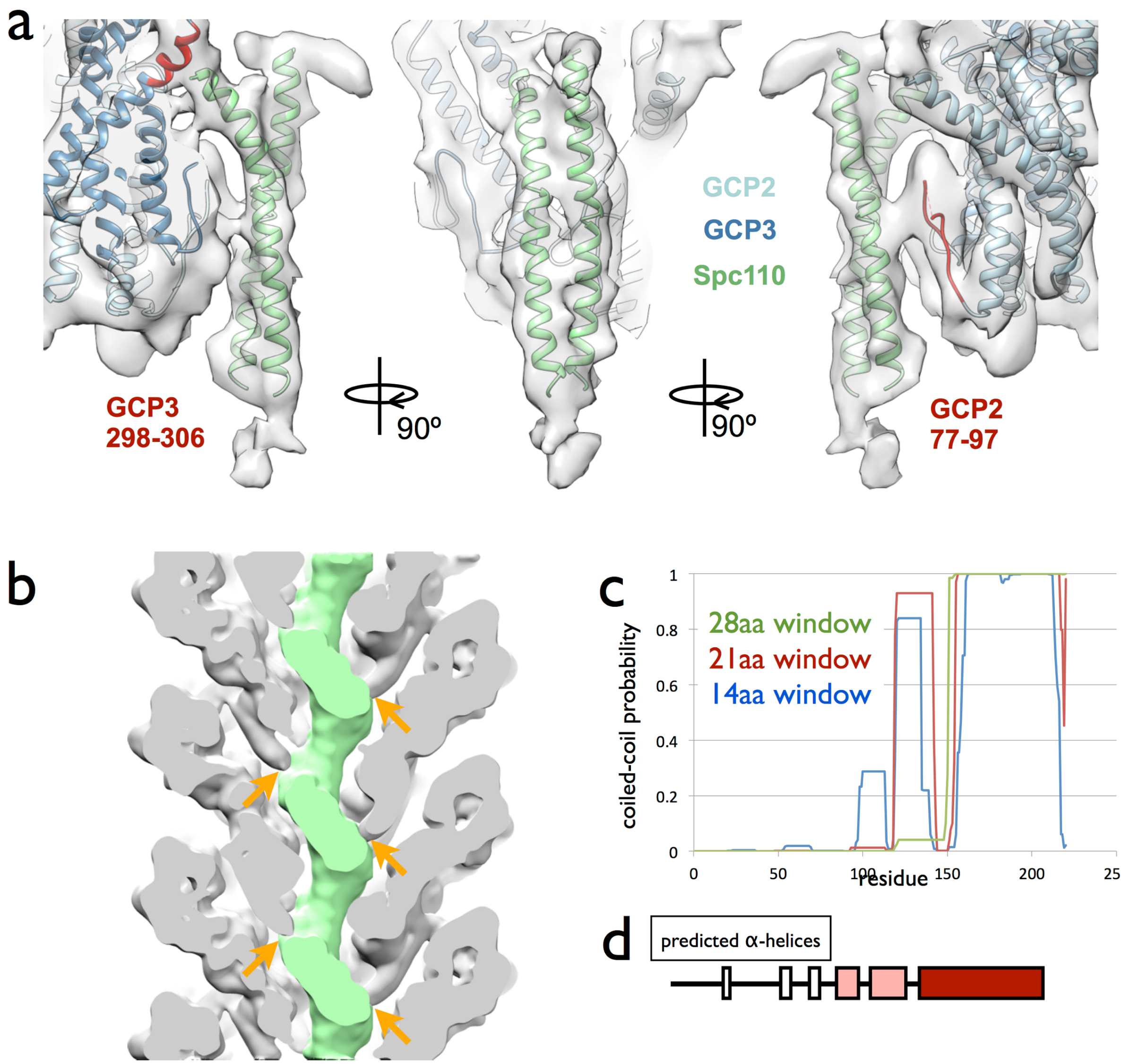


Figure S6. Spc110 coiled-coil interacts with N-terminal domains of GCP2 and GCP3. a) The density predicted to correspond to part of Spc110$^{1\text{-}220}$ resembles a two-stranded coiled-coil. A generic two-stranded coiled coil was used to model Spc110, and flexibly fit into the EM density. The closest contacts with Spc110 are between residues 77-97 of GCP2 (which includes 9 residues not included in our model) near the very base of γTuSC, and further up the structure with residues 119-126 of GCP3. b) A central slice of the γTuSC^{CC} filament in the closed state, low-pass filtered at 15 Å. The γTuSC and modeled Spc110 structure is colored grey, and disordered density that runs along the center of the filament is colored green. Arrows denote the connections between the ends of the modeled Spc110 structure and the disordered density, which suggest that the disordered density is composed of flexible regions of Spc110$^{1\text{-}220}$. c) Predicted propensity for coiled-coil formation of the Spc110$^{1\text{-}220}$ primary sequence, calculated over windows of 14, 21, or 28 amino acids. d) Secondary structure prediction of Spc110$^{1\text{-}220}$, with α-helices denoted as boxes. Pink boxes have a weaker predicted coiled-coil probability, and red boxes a high probability.

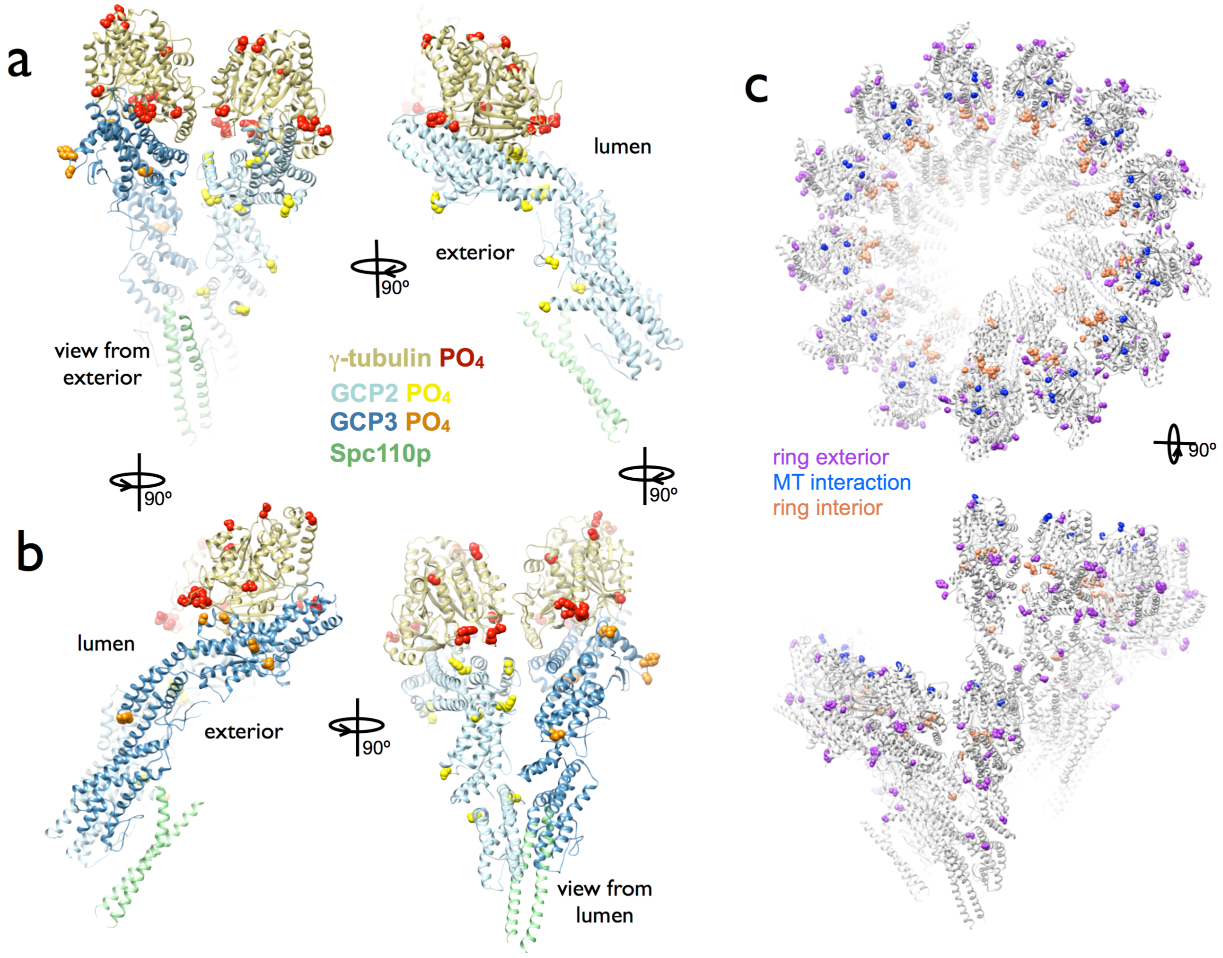


Figure S7. Mapping phosphorylation sites on the γTuSC structure. a) A single γTuSC is shown with residues known to be phosphorylated highlighted as spheres in red, yellow, and orange for sites on γ-tubulin, GCP2, and GCP3 respectively. c) The model of yeast γTuRC with known phosphorylation sites rendered as spheres. Sites freely accessible on the outer surface of the ring are colored purple, sites at the plus end of γ-tubulin predicted to be involved in longitudinal contacts with the microtubule in blue, and sites on the ring interior that would be inaccessible when a microtubule is bound are coral.

REFERENCES

1 Choy, R. M., Kollman, J. M., Zelter, A., Davis, T. N. & Agard, D. A. Localization and orientation of the gamma-tubulin small complex components using protein tags as labels for single particle EM. *Journal of structural biology* **168**, 571-574, doi:10.1016/j.jsb.2009.08.012 (2009).

2 Kollman, J. M. *et al.* The structure of the gamma-tubulin small complex: implications of its architecture and flexibility for microtubule nucleation. *Molecular biology of the cell* **19**, 207-215, doi:10.1091/mbc.E07-09-0879 (2008).

3 Oegema, K. *et al.* Characterization of two related Drosophila gamma-tubulin complexes that differ in their ability to nucleate microtubules. *J Cell Biol* **144**, 721-733 (1999).

4 Zheng, Y., Wong, M. L., Alberts, B. & Mitchison, T. Nucleation of microtubule assembly by a gamma-tubulin-containing ring complex. *Nature* **378**, 578-583 (1995).

5 Kollman, J. M., Merdes, A., Mourey, L. & Agard, D. A. Microtubule nucleation by gamma-tubulin complexes. *Nature reviews. Molecular cell biology* **12**, 709-721, doi:10.1038/nrm3209 (2011).

6 Moritz, M., Braunfeld, M. B., Guenebaut, V., Heuser, J. & Agard, D. A. Structure of the gamma-tubulin ring complex: a template for microtubule nucleation. *Nat Cell Biol* **2**, 365-370 (2000).

7 Oakley, B. R., Oakley, C. E., Yoon, Y. & Jung, M. K. Gamma-tubulin is a component of the spindle pole body that is essential for microtubule function in Aspergillus nidulans. *Cell* **61**, 1289-1301, doi:0092-8674(90)90693-9 [pii] (1990).

8 Kilmartin, J. V. & Goh, P. Y. Spc110p: assembly properties and role in the connection of nuclear microtubules to the yeast spindle pole body. *The EMBO journal* **15**, 4592-4602 (1996).

9 Sundberg, H. A. & Davis, T. N. A mutational analysis identifies three functional regions of the spindle pole component Spc110p in Saccharomyces cerevisiae. *Molecular biology of the cell* **8**, 2575-2590 (1997).

10 Kollman, J. M., Polka, J. K., Zelter, A., Davis, T. N. & Agard, D. A. Microtubule nucleating gamma-TuSC assembles structures with 13-fold microtubule-like symmetry. *Nature* **466**, 879-882, doi:10.1038/nature09207 (2010).

11 Tilney, L. G. *et al.* Microtubules: evidence for 13 protofilaments. *J Cell Biol* **59**, 267-275 (1973).

12 O'Toole, E. T., Winey, M. & McIntosh, J. R. High-voltage electron tomography of spindle pole bodies and early mitotic spindles in the yeast Saccharomyces cerevisiae. *Molecular biology of the cell* **10**, 2017-2031 (1999).

13 Erlemann, S. *et al.* An extended gamma-tubulin ring functions as a stable platform in microtubule nucleation. *J Cell Biol* **197**, 59-74, doi:10.1083/jcb.201111123 (2012).

14 Sui, H. & Downing, K. H. Structural basis of interprotofilament interaction and lateral deformation of microtubules. *Structure* **18**, 1022-1031, doi:10.1016/j.str.2010.05.010 (2010).

15 Keating, T. J. & Borisy, G. G. Immunostructural evidence for the template mechanism of microtubule nucleation. *Nat Cell Biol* **2**, 352-357 (2000).

16 Wiese, C. & Zheng, Y. A new function for the gamma-tubulin ring complex as a microtubule minus-end cap. *Nat Cell Biol* **2**, 358-364 (2000).

17 Aldaz, H., Rice, L. M., Stearns, T. & Agard, D. A. Insights into microtubule nucleation from the crystal structure of human gamma-tubulin. *Nature* **435**, 523-527 (2005).

18 Rice, L. M., Montabana, E. A. & Agard, D. A. The lattice as allosteric effector: structural studies of alphabeta- and gamma-tubulin clarify the role of GTP in microtubule assembly. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 5378-5383, doi:10.1073/pnas.0801155105 (2008).

19 Egelman, E. H. The iterative helical real space reconstruction method: surmounting the problems posed by real polymers. *Journal of structural biology* **157**, 83-94, doi:10.1016/j.jsb.2006.05.015 (2007).

20 Johnson, V., Ayaz, P., Huddleston, P. & Rice, L. M. Design, overexpression, and purification of polymerization-blocked yeast alphabeta-tubulin mutants. *Biochemistry* **50**, 8636-8644, doi:10.1021/bi2005174 (2011).

21 Vinh, D. B., Kern, J. W., Hancock, W. O., Howard, J. & Davis, T. N. Reconstitution and characterization of budding yeast gamma-tubulin complex. *Molecular biology of the cell* **13**, 1144-1157 (2002).

22 Choi, Y. K., Liu, P., Sze, S. K., Dai, C. & Qi, R. Z. CDK5RAP2 stimulates microtubule nucleation by the gamma-tubulin ring complex. *J Cell Biol* **191**, 1089-1095 (2010).

23 Lowe, J., Li, H., Downing, K. H. & Nogales, E. Refined structure of alpha beta-tubulin at 3.5 A resolution. *J Mol Biol* **313**, 1045-1057, doi:10.1006/jmbi.2001.5077 (2001).

24 Ravelli, R. B. *et al.* Insight into tubulin regulation from a complex with colchicine and a stathmin-like domain. *Nature* **428**, 198-202, doi:10.1038/nature02393 (2004).

25 Holt, L. J. *et al.* Global analysis of Cdk1 substrate phosphorylation sites provides insights into evolution. *Science* **325**, 1682-1686, doi:10.1126/science.1172867 (2009).

26 Albuquerque, C. P. *et al.* A multidimensional chromatography technology for in-depth phosphoproteome analysis. *Molecular & cellular proteomics : MCP* **7**, 1389-1396, doi:10.1074/mcp.M700468-MCP200 (2008).

27 Holinger, E. P. *et al.* Budding yeast centrosome duplication requires stabilization of Spc29 via Mps1-mediated phosphorylation. *J Biol Chem* **284**, 12949-12955, doi:10.1074/jbc.M900088200 (2009).

28 Keck, J. M. *et al.* A cell cycle phosphoproteome of the yeast centrosome. *Science* **332**, 1557-1561 (2011).

29 Lin, T. C. *et al.* Phosphorylation of the yeast gamma-tubulin Tub4 regulates microtubule function. *PLoS One* **6**, e19700, doi:10.1371/journal.pone.0019700 (2011).

30 Vogel, J. *et al.* Phosphorylation of gamma-tubulin regulates microtubule organization in budding yeast. *Dev Cell* **1**, 621-631 (2001).

31 Nazarova, E. *et al.* Distinct roles for antiparallel microtubule pairing and overlap during early spindle assembly. *Molecular biology of the cell* **24**, 3238-3250, doi:10.1091/mbc.E13-05-0232 (2013).

32 Samejima, I., Miller, V. J., Groocock, L. M. & Sawin, K. E. Two distinct regions of Mto1 are required for normal microtubule nucleation and efficient association with the gamma-tubulin complex in vivo. *J Cell Sci* **121**, 3971-3980, doi:10.1242/jcs.038414 (2008).

33 Goshima, G., Mayer, M., Zhang, N., Stuurman, N. & Vale, R. D. Augmin: a protein complex required for centrosome-independent microtubule generation within the spindle. *J Cell Biol* **181**, 421-429, doi:10.1083/jcb.200711053 (2008).

34 Rout, M. P. & Kilmartin, J. V. Yeast spindle pole body components. *Cold Spring Harbor Symp. Quant. Biol.* **56**, 687-692 (1991).

35 Donaldson, A. D. & Kilmartin, J. V. Spc42p: a phosphorylated component of the S. cerevisiae spindle pole body (SPD) with an essential function during SPB duplication. *J. Cell Biol.* **132**, 887-901 (1996).

36 Zheng, Q. S., Braunfeld, M. B., Sedat, J. W. & Agard, D. A. An improved strategy for automated electron microscopic tomography. *J. Struct. Biol.* **147**, 91-101, doi:10.1016/j.jsb.2004.02.005 (2004).

37 Scheres, S. H. W., Melero, R., Valle, M. & Carazo, J.-M. Averaging of electron subtomograms and random conical tilt reconstructions through likelihood optimization. *Structure (London, England : 1993)* **17**, 1563-1572, doi:10.1016/j.str.2009.10.009 (2009).

38 Förster, F., Pruggnaller, S., Seybert, A. & Frangakis, A. S. Classification of cryo-electron sub-tomograms using constrained correlation. *J. Struct. Biol.* **161**, 276-286, doi:10.1016/j.jsb.2007.07.006 (2008).

39 Sikorski, R. S. & Hieter, P. A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in Saccharomyces cerevisiae. *Genetics* **122**, 19-27 (1989).

40 Ohi, M., Li, Y., Cheng, Y. & Walz, T. Negative Staining and Image Classification - Powerful Tools in Modern Electron Microscopy. *Biol Proced Online* **6**, 23-34 (2004).

41 Quispe, J. *et al.* An improved holey carbon film for cryo-electron microscopy. *Microsc Microanal* **13**, 365-371, doi:S1431927607070791 [pii]
10.1017/S1431927607070791 (2007).

42 Mindell, J. A. & Grigorieff, N. Accurate determination of local defocus and specimen tilt in electron microscopy. *J Struct Biol* **142**, 334-347, doi:S1047847703000698 [pii] (2003).

43 Egelman, E. H. A robust algorithm for the reconstruction of helical filaments using single-particle methods. *Ultramicroscopy* **85**, 225-234, doi:S0304399100000620 [pii] (2000).

44 Sachse, C. *et al.* High-resolution electron microscopy of helical specimens: a fresh look at tobacco mosaic virus. *J Mol Biol* **371**, 812-835, doi:S0022-2836(07)00759-0 [pii]
10.1016/j.jmb.2007.05.088 (2007).

45 Frank, J. *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*. (Academic Press, Inc., 1996).

46 Pettersen, E. F. *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605-1612 (2004).

47 Muller, E. G. A glutathione reductase mutant of yeast accumulates high levels of oxidized glutathione and requires thioredoxin for growth. *Molecular biology of the cell* **7**, 1805-1813 (1996).

48 Muller, E. G. *et al.* The organization of the core proteins of the yeast spindle pole body. *Molecular biology of the cell* **16**, 3341-3352, doi:10.1091/mbc.E05-03-0214 (2005).

49 Shimogawa, M. M., Widlund, P. O., Riffle, M., Ess, M. & Davis, T. N. Bir1 is required for the tension checkpoint. *Molecular biology of the cell* **20**, 915-923, doi:10.1091/mbc.E08-07-0723 (2009).

50 Pei, J., Kim, B. H. & Grishin, N. V. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic acids research* **36**, 2295-2300, doi:10.1093/nar/gkn072 (2008).

51 Sali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234**, 779-815, doi:10.1006/jmbi.1993.1626 (1993).

52 Trabuco, L. G., Villa, E., Mitra, K., Frank, J. & Schulten, K. Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure* **16**, 673-683, doi:10.1016/j.str.2008.03.005 (2008).

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

____
Author Signature

Date September 5, 2016