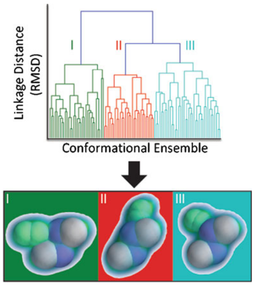
J. Am. Soc. Mass Spectrom. (2017) 28:1991–2000
DOI: 10.1007/s13361-017-1757-1

CRITICAL INSIGHT

Coming to Grips with Ambiguity: Ion Mobility-Mass Spectrometry for Protein Quaternary Structure Assignment

Joseph D. Eschweiler, Aaron T. Frank, Brandon T. Ruotolo

Department of Chemistry, University of Michigan, Ann Arbor, MI 48109, USA



Abstract. Multiprotein complexes are central to our understanding of cellular biology, as they play critical roles in nearly every biological process. Despite many impressive advances associated with structural characterization techniques, large and highly-dynamic protein complexes are too often refractory to analysis by conventional, high-resolution approaches. To fill this gap, ion mobility-mass spectrometry (IM-MS) methods have emerged as a promising approach for characterizing the structures of challenging assemblies due in large part to the ability of these methods to characterize the composition, connectivity, and topology of large, labile complexes. In this Critical Insight, we present a series of bioinformatics studies aimed at assessing the information content of IM-MS datasets for building models of multiprotein structure. Our computational data highlights the limits of current coarse-graining approaches, and compelled us to develop an improved workflow for multiprotein topology modeling, which we benchmark against a subset of the multiprotein complexes within the PDB. This improved workflow has allowed us to ascertain both the minimal experimental restraint sets required for generation of high-confidence multiprotein topologies, and quantify the ambiguity in models where insufficient IM-MS information is available. We conclude by projecting the future of IM-MS in the context of protein quaternary structure assignment, where we predict that a more complete knowledge of the ultimate information content and ambiguity within such models will undoubtedly lead to applications for a broader array of challenging biomolecular assemblies.

Keywords: Native mass spectrometry, Structural proteomics, Bioinformatics, Molecular dynamics simulations, Protein network

Received: 18 April 2017/Revised: 4 July 2017/Accepted: 5 July 2017/Published Online: 27 July 2017

Introduction

Structural characterization of the multicomponent complexes that form the functional units of the "interactome," specifically protein complexes, represents a major challenge for structural biology [1, 2]. Owing to their large size, low copy numbers, and intrinsic heterogeneity and lability, important targets are too often refractory to analysis by traditional techniques such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, or electron microscopy, despite impressive advances in these fields [3, 4]. Alternative approaches for characterizing difficult multicomponent structures may result in low-resolution or sparse datasets, such as those generated from small-angle scattering [5] or covalent labeling/crosslinking methodologies [6]. Circumventing the limitations of a single technique, integration of datasets from multiple experiments has been shown to be a potent approach for characterizing multiprotein complexes [7] as often times these datasets provide complementary information. This workflow, commonly referred to as integrative structural biology, has progressed rapidly due largely to advances in computational techniques that have made it possible to encode different types of experimental datasets as spatial restraints in a single modeling workflow [8].

Generally, an integrative modeling workflow is an iterative process described by four major steps: (1) the gathering of experimental data, (2) the translation of such data into spatial restraints, (3) the generation of an ensemble of putative structures that satisfy the experimentally defined restraints, and (4) the characterization of the ensembles generated, where ambiguities are identified and used to refine structural hypotheses. This process may then be iterated as necessary in order to

Electronic supplementary material The online version of this article (doi:10.1007/s13361-017-1757-1) contains supplementary material, which is available to authorized users.

Correspondence to: Brandon Ruotolo; *e-mail:* bruotolo@umich.edu

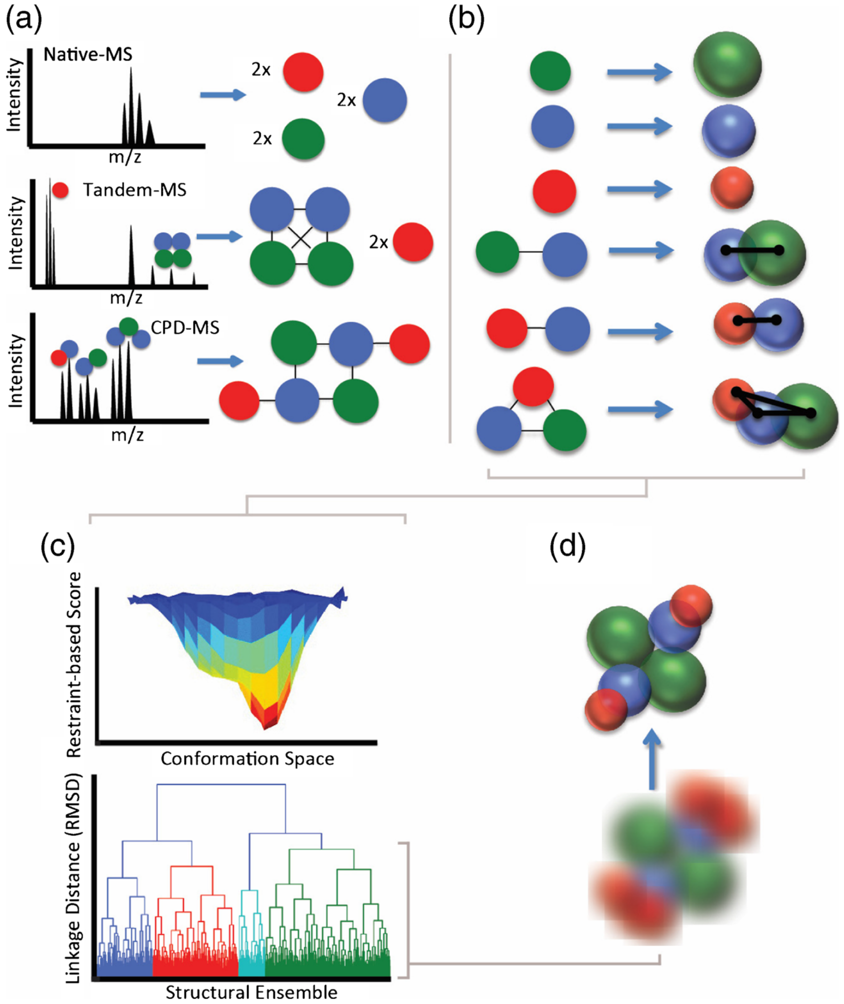


Figure 1. A general workflow for IM-MS-based modeling. (**a**) Native-MS, tandem-MS, and solution-phase disruption-MS yield increasing amounts of composition and connectivity information for a multiprotein complex. This information can be encoded with varying levels of ambiguity based on the information available. (**b**) IM-MS data can be included to build a 3D topology model. Individual subunits or domains can be encoded as spheres with radius derived from their measured CCS, whereas exact distances between subunits can be derived from CCS measurements of dimeric and trimeric species. (**c**) Optimization of the experimentally defined scoring function using a Monte Carlo method provides unbiased sampling of potential structures for high-stoichiometry complexes. These structures form an ensemble that is subjected to clustering analysis to mine for predominant structural families. (**d**) Structural families detected by clustering can be characterized in aggregate using kernel density functions, mean structures, and standard deviations, or individual structures can be identified as representative of the family

resolve ambiguities to the extent allowed by the experimental restraints utilized [8]. MS-based methods such as chemical crosslinking [9–11], native-MS [12, 13], and ion mobility-MS [14] have garnered much attention as valuable experiments within such integrative structural biology frameworks. Of these MS-based technologies, ion-mobility-mass spectrometry (IM-MS) is uniquely positioned for the interrogation of multiprotein structure [15]. Unlike solution-phase measurements, which may report on the average of an ensemble of proteoforms, conformers, or oligomerization states, IM-MS datasets can be used to discern the relative proportions of these species within mixtures, and interrogate their composition, connectivity, and collision cross-sections (CCSs) individually [16]. Due to its unique capabilities in protein structure analysis, IM-MS is often deployed to determine coarse-grained (CG) protein topology models for assemblies that have resisted previous characterization attempts, often in combination with other forms of biophysical data.[17, 18].

Figure 1a illustrates the potential information content often derived from native MS datasets. While direct analysis of the masses of intact complexes can often provide unambiguous information about the protein composition and stoichiometry [19], it is also useful to interrogate solution or gas-phase disassembly products to further elucidate connectivity and structural modularity. To this end, methods for solution [20, 21] and gas-phase [4–22] disruption of multiprotein complexes are actively being developed to increase the number of observable sub-complexes and, therefore, the overall information content of the experiment. In addition to the composition and connectivity information garnered by MS, IM-MS (Figure 1b) provides 3D structural information on both monomeric and oligomeric protein ions in the form of CCSs [23, 24]. Since multiple methods are available for the accurate calculation of CCS values from in silico models [25–27], it is possible to assign putative structures to the signals observed in the IM-MS experiment.

Despite being used to restrain rigorous dynamics experiments for peptides [28] and small proteins [29] for decades, our ability to extract structural information from CCS measurements of large proteins and multiprotein complexes is still evolving. A recent comprehensive analysis of the PDB revealed that the general amount of CCS variance exhibited by proteins increases for high mass and stoichiometry protein complexes, indicating increased CCS information content is available for such assemblies [26]. These observations corroborate earlier experimental results showing that the oligomerization patterns of homomeric protein complexes can be discerned in many cases based on CCS trends as a function of complex stoichiometry [30].

Methods for extracting topological information for large, heteromeric protein complexes are, however, less developed. Early procedures for optimizing pairwise and trimeric subunit interactions were based on a linear search for conformations, using spherical subunit representations that satisfied experimental CCS restraints [31]. Although the spherical representation of protein subunits possesses obvious limitations when modeling highly aspherical subunits such as multidomain proteins, spheres still represent the primary component in IM-MS-based modeling because of their trivial geometric relationship to the CCS parameter, their ease of implementation in computational workflows, and their facile relationship to protein–protein interaction geometries. Subsequently described IM-MS workflows aimed at the generation of protein quaternary structure models (Figure 1c) utilized a Monte Carlo approach for sampling orientations of spheres that satisfied excluded volume, symmetry, connectivity, and CCS restraints in order to yield an ensemble of structures that can be interrogated via hierarchical clustering methodologies [31–33]. Such IM-MS-derived models have been favorably compared with structures produced using more mature experimental workflows, indicating a promising level of cooperativity between CCS measurements and other biophysical parameters commonly used in protein complex model generation [32]. This general approach has been used to elucidate the topological features of the DNA replisome [31, 34], ribosomal initiation factor complexes [35], and ATPases [36], all providing critical structural insights as well as methodological enhancements. More recently, surface induced dissociation (SID) coupled to IM-MS and covalent labeling has been applied to build a complete model of the toyocamycin nitrile hydratase complex [37] by leveraging the sub-complexes produced both through controlled disruption in solution and SID.

Despite these promising examples, many questions remain about the ability to unambiguously assign protein topology using IM-MS datasets (Figure 1d). Most of these questions surround the potential errors introduced when high levels of coarse-graining are applied, the interpretation of structural ensembles generated from IM-MS modeling approaches, and the confidence levels associated with IM-MS structures in a general sense [38]. Additionally, questions remain regarding the extent of structural rearrangement apparent in some proteins and complexes in the gas phase, a topic that has been investigated in detail elsewhere [33, 39]. In this Critical Insight, we seek to critically evaluate the information content of IM-MS for protein quaternary structure assignment in cases where we can assume a strong memory of solution-phase structure. Based on many of the challenges described above, we develop a new generalized algorithm for translating IM-MS datasets into structural models and benchmark our new method against many known topologies present in the PDB. We continue by quantifying, for the first time, the ambiguity present in under-restrained models, and suggest approaches for mitigating such effects. We conclude by projecting the future of IM-MS derived models of protein quaternary structure.

Assessing Coarse-Graining Errors in Multiprotein Models Generated from IM-MS Data

In workflows that utilize IM-MS data to restrain models of protein quaternary structure, it is typically assumed that the protein components of the assembly can be accurately represented by spheres defined by either their measured or estimated CCS. Although many reports have demonstrated a strong correlation between experimental CCS measurements and CCS values extracted from solution-phase protein models, the strength of this correlation can depend on the domain structure and globularity of the protein analyte in question [40, 41]. Moreover, the magnitude and nature of the errors incorporated into IM-MS multiprotein models through the coarse-graining process are currently unknown. In order to investigate such coarse-graining errors, we extracted a non-redundant set of 191 high-resolution protein complex structures from the 3D complex set database [42], and developed a method for the rapid generation of CG structures based on these entries where the extent of coarse-graining can be treated as a variable. The first step in our protocol involves extracting coordinates and center-of-mass values for each subunit within the protein complex. Next, the CCS values are calculated for each subunit using the projection approximation function within the IMPACT library [26]. To generate the initial CG model at subunit resolution, we placed spheres having radii corresponding to the projected area of the subunits at the center of mass for each subunit in the complex. To evaluate the model, the projected area of the high-resolution structure was compared with that calculated from the CG model.

Our results suggest that a significant number of the protein complexes currently available within the PDB contain subunits that are not accurately represented when subunit-level coarse-graining is applied. As shown in Figure 2a–c, subunit-level coarse-graining very often results in large deviations in CCS compared with the reference. We used a 5% deviation in the CCS values obtained for CG models when compared to reference CCSs for the corresponding all-atom reference structure to define a 'significant' error threshold in our analysis, as such defects reflect, in our view, both the maximum error that can be introduced into a model before losing significant topology information as well as the maximum error value carried by experimental restraint information recovered for large protein complexes by IM-MS [15]. Specifically, over 38% of the protein complexes studied here contained significant errors (greater than 5%) when this level of coarse-graining was applied. We also note that the error distribution associated with this level of coarse-graining is highly asymmetric, containing many structures having CG errors greater than 10%.

A more detailed analysis of the structures within the survey reveals that proteins with multiple domains are most susceptible to high CG errors, especially those proteins having domains connected by long linker regions. Interestingly, however, we found no correlation between the CCS/mass ratio of individual subunits and their propensity to introduce error into the model, indicating that the overall packing density of the protein does not play a major role in the CG errors on display in Figure 2d and e. Based on this data, we hypothesized that coarse-graining at the domain level should eliminate the majority of the errors we observed from our subunit-resolution CG modeling experiments. To investigate this, we implemented a k-means clustering method [43] in SciPy [44] to heuristically detect protein domain structure over a range of thresholds associated with protein and domain mass (See Supporting Information for details). The results associated with these higher-resolution CG structures are shown in Figure 2f and g, and reveal a strong relationship between the resolution of the CG structures and the propensity for CG error we record during our analysis. Figure 2g, for example, shows that the fraction of protein complexes with significant errors drops to ~2% when domain-level CG is applied to the same pool of structures analyzed in Figure 2e. We find that error-prone structures that persist in our analysis are largely those containing extremely long linker regions or aspherical domains that remain inaccurately represented using domain-level coarse-graining.

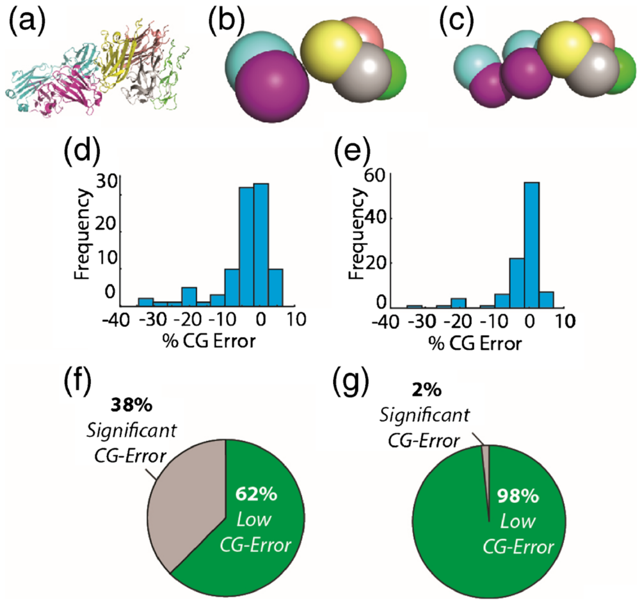


Figure 2. Coarse-graining error for domain and subunit-level representations. (**a**) An example high resolution reference structure PDB ID 4MXW with subunits color coded. (**b**) A coarse-grained model of 4MXW at the subunit level. (**c**) A coarse-grained model of 4MXW at the domain level. (**d**) One hundred ninety-one non-redundant protein topologies were coarse-grained at the subunit-level. The coarse-graining error distribution for this level of coarse-graining is shown, with bin sizes of approximately 4% error along the X-axis. (**e**) The coarse-graining error distribution for the same set of protein topologies coarse-grained at the domain level. (**f**) Subunit-level coarse-graining introduced significant CCS errors for 38% of the complexes in our set. (**g**) When coarse-grained at the domain level, only 2% of topologies had significant coarse-graining errors introduced

Benchmarking the Information Content of IM-MS Datasets for Modeling Known Protein Complexes

To generate ensembles of putative structures based on IM-MS-derived data, we developed a program for the interpretation and optimization of diverse MS and/or IM-derived restraint sets. This program, referred to as IMMS_modeler, was built using connectivity and distance restraints from the Integrated Modeling Platform (IMP) [8], some of which were implemented previously [32]. Novel aspects of our approach include: (1) the use of a restraint file for facile input of new data, (2) the ability to use new mathematical definitions within the scoring function, and (3) a new Monte Carlo algorithm that enables a significantly broader sampling restraint space. By default, IMMS_modeler generates ensembles of 1000 structures that satisfy all of the declared restraints. We found this amount of structures to be a representative sample of structural space for most complexes, and have based the following experiments on these ensembles. All CCS calculations were performed off-line using the projected area function in the IMPACT library (see Supporting Information for details).

In order to thoroughly evaluate our method as a general approach for modeling multiprotein complexes, we set out to benchmark IMMS_modeler against known protein complex topologies with varying levels of restraint information. In these experiments, we generated CG models at the resolution of individual protein subunits for a small subset of complex topologies used in the previous experiment. For simplicity, we focused this stage of our analysis only on those protein complexes that did not show significant CG error, as described in the above section (Figure 2). Despite these limitations, the geometric principles described here are transferrable to models created at higher levels of CG resolution.

On the Positive Predictive Power of IM-MS datasets

In order to characterize the information content associated with CCS measurements of intact protein complexes and subcomplexes when used to define inter-protein distances and geometry in the context of a search of potential quaternary structures (which we define as 'internal restraints'), we simulated IM-MS datasets for at least five non-redundant complex topologies for protein trimers, tetramers, pentamers, and hexamers (Figure 3). Although some of the complexes used to generate the analysis shown in Figure 3 contain symmetric elements, no symmetry restraints were implemented to avoid bias. All restraint sets contained detailed information regarding the connectivity of the complex, as well as the CCS of the intact assembly, as in our view these restraints are essential for any IM-MS based quaternary structure assignment. In addition to this information, restraint sets containing varying numbers of the 'internal restraints' described above, which correspond to the pair-wise distance restraints that are commonly obtained from native IM-MS datasets, are also included in our analyses [21, 30]. We note that although 3D systems are completely restrained by a minimum of 3N-6 fixed distance restraints (where N is number of bodies), the restraints used in this report attempt to simulate real IM-MS data. Specifically, restraints mined from IM-MS data contain error, often producing predictive values that are less than those generated through precise distance geometries. For purposes of this analysis, the structures generated using our method were defined as true positives (native-like topologies) if they had RMSD values of less than 5 Å relative to the reference structure, and we defined a positive predictive value (PPV) as the fraction of true positive structures within the ensemble of structures sampled using a given restraint set (see Supplementary Figure S2 for examples of structures filtered out of these analyses using RMSD).

As expected, our results reveal a positive relationship between the number of internal CCS restraints available for a complex and the positive predictive value for a given modeling effort. For trimeric protein complexes (Figure 3a), the ensemble is enriched for true positives with the addition of internal distance restraints between subunits. Here, due to the trivial relationship between the CCS and the angle of subunits within the complex, the model should be fully restrained by the global CCS plus any two IM-derived distance restraints [31]. Notably, there is one outlier structure that seemingly refutes this general conclusion; however, our analysis also suggests that the global CCS restraint becomes less sensitive when large disparities exist in the CCS of each component, allowing us to rationalize all of the results shown (Supplementary Figure S3). Higher stoichiometry complexes (Figure 3b–d), exhibit similarly strong increases in PPV in a manner correlated with the number of internal restraints included. We note that the number of restraints necessary to reach a PPV > 0.8, where 80% of the structures identified in the ensemble are within 5 Å of the 'true' structure, increases rapidly as the number of subunit increases, further

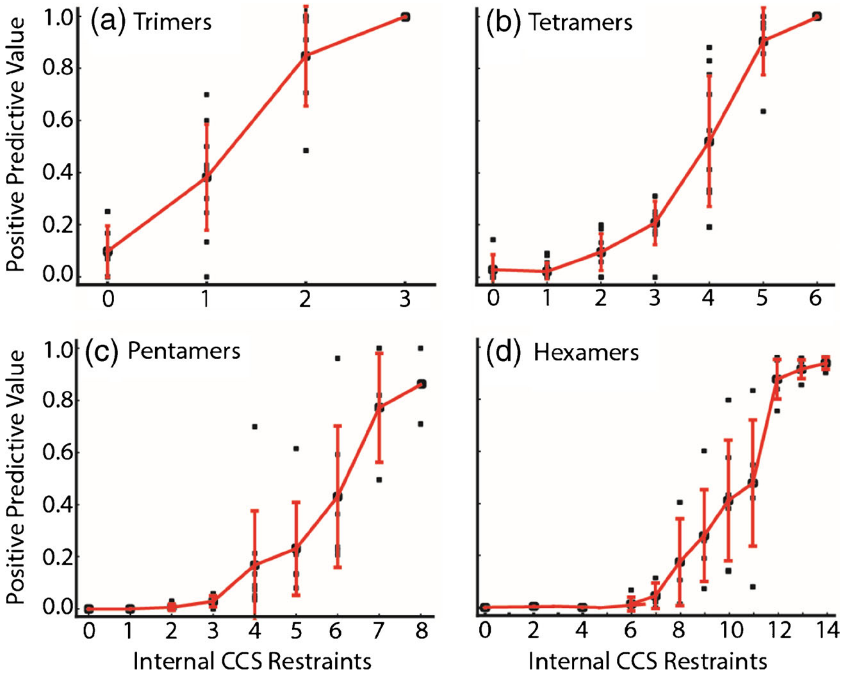


Figure 3. Positive predictive values of the IM-MS restraint sets plotted as a function of the number of internal CCS-derived restraints. At least five non-redundant topologies from the PDB were considered for each number of subunits, (**a**) trimers, (**b**) tetramers, (**c**) pentamers, and (**d**) hexamers. Each restraint set was manually curated to ensure the data reflected data that could be reasonably generated through existing IM-MS technologies

motivating the need to develop new methods and technologies for the comprehensive generation of native-like subcomplexes for IM-MS analysis [20, 45, 46].

Characterizing Ambiguity in the Structural Ensembles Defined by IM-MS

Although the PPV is a valuable metric for comparing the information content of multiple restraint sets, interpretation of PPV values for individual datasets can be challenging. This is due to the fact that members of a structural ensemble generated by the IMMS-modeling approach described here are not randomly distributed, and in many cases can be clustered into distinct sub-distributions, or structural families. Pairwise relationships between structures within an ensemble can be described by a pairwise RMSD matrix, which can in turn be interrogated using hierarchical clustering to determine groups of highly related structures. Alternatively, other similarity measures can be implemented to describe structural relationships between models, including the ultrafast similarity score [47], or distance matrix RMSD [48], which each may have its own advantages depending on the geometries present in the ensemble. For the computational data described in this Critical Insight, a detailed analysis of the structural ensemble produced from an IM-MS restrained search of protein topology space regularly reveals useful information, in addition to what is provided by the PPV value analysis shown in Figure 3 alone. In the sections below, we discuss the interpretation of hierarchical clustering datasets in the context of such IM-MS restrained models, focusing on our recent efforts to define and quantify the ambiguity and resolution within the IM-MS data.

A hierarchical clustering dendrogram (as shown in Figure 4) illustrates the relationship between all structures within an ensemble. The number of clusters depends on the 'cut point' chosen during dendrogram analysis, a value that is typically a user defined parameter. For example, our algorithm automatically defaults to a dendrogram cut point that generates clusters at linkages that exhibit greater than 70% of the maximum RMSD in the entire matrix analyzed. Our ensemble analysis workflow evaluates the in-cluster RMSD as it compares to the average RMSD of the ensemble, as well as the cross-cluster RMSD, revealing distinct structural families that define the identified clusters (Figure 4). It is worth noting that the application of IM-MS restraints often leads to the type of model ambiguity shown in Figure 4 for large hetero-protein targets [32]. Indeed, such ambiguity may, in some cases, represent the native ensemble of protein complex structures associated with function [36, 49]. Commonly, however, such uncertainty is due to incomplete structural information and can be resolved through the application of additional restraints [18, 50] (see below for examples).

As mentioned above, the in-cluster RMSD can be a valuable metric for quantitatively expressing the ambiguity within a cluster. However, when evaluating biomolecular structures, qualitative and visual expression of ambiguity is often more facile to interpret. In order to fill this gap for IM-MS-derived models, we developed a new method for visualizing the ambiguity within a structural family using kernel density functions [44, 51]. In this method, the coordinates within a structural family or ensemble are aligned, and each subunit coordinate is uniformly populated with protein density as a sphere corresponding to its collision cross-section. Next, the Gaussian kernel function is estimated for this volume of coordinates, and then visualized. For the workflow described here, we utilize the Mayavi Library [52] in Python to visualize the kernel densities. As illustrated in Figure 4c, this kernel density function approach allows for the visualization of structural ambiguity present within an ensemble, information that is likely vital for the detailed interpretation of structural ensembles defined by sparse sets of restraints.

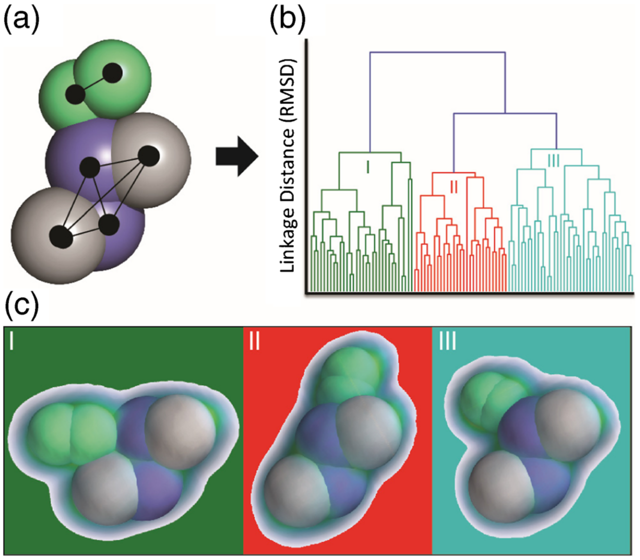


Figure 4. Parsing structural ensembles generated with ambiguous restraint sets. (a) A restraint set was generated for 2AFH, a nitrogenase heterotetramer (purple and grey) bound to the dimeric nucleotide switch protein (green). The binding location and pose for the nucleotide switch protein is not defined in the restraint set and a CCS-filtered structural ensemble contains many putative structures. (b) Hierarchical clustering of the ensemble reveals three distinct structural families within the ensemble, greatly simplifying the analysis. (c) Plotting the kernel density function of each structural family reveals high resolution within all families

Leveraging Symmetry and Modularity to Resolve Ambiguity within IM-MS Model Ensembles

To further evaluate IM-MS-based quaternary structure assignments in a general sense, as well as the newly developed methods described here, we chose two case studies that illustrate real-world examples of challenging modeling targets. As shown in Figure 3, the number of restraints needed to accurately recapitulate the topology of a multiprotein complex increases linearly, creating challenges for integrative modeling of these complexes. However, in the data shown below, we demonstrate that by leveraging the modularity and symmetry within high-stoichiometry complexes, it is possible to circumvent these limitations.

As an example of a symmetry restraint applied in order to resolve ambiguity within an IM-MS restrained ensemble of protein quaternary structures, we built models of the Large T-antigen (LTag) complex bound to p53. LTag is a hexameric ring structure that binds p53 monomers in a stoichiometric and symmetric fashion around the ring [53]. Assuming a comprehensive protein–protein connectivity dataset from Native MS, we searched for a minimal IM-MS restraint set to recapitulate the known topology of LTag-p53 with C6 symmetry. Our first attempt utilized only connectivity and global CCS information to generate a structural ensemble. (Figure 5a). For this ensemble, we observe three structural families, with relatively little resolution between them. Each family is represented by a very broad distribution of RMSD values relative to the reference structure, indicating that both the accuracy and effective resolution of the structural models created in this search are low. The kernel density function estimated for each structural family also illustrates the poor resolution generated from this restraint set.

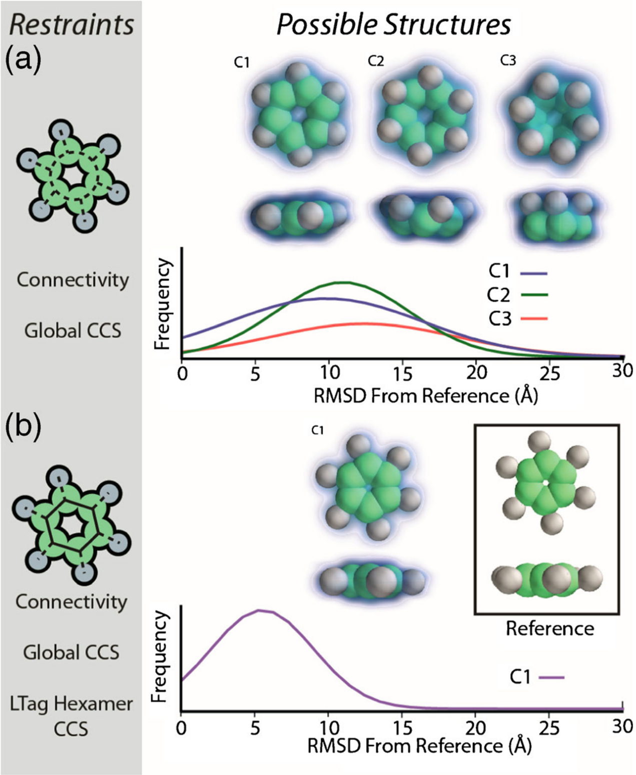


Figure 5. Modeling the topology of hexameric LTag bound to p53 using the symmetry restraint. Two restraint sets (left panels), (**a**) and (**b**) were used to generate structural ensembles that were evaluated using hierarchical clustering, kernel density functions, and RMSD distributions (right panels)

In order to resolve the above ambiguity, we utilized restraints associated with the CCS of the LTag hexamer in addition to the intact complex CCS and overall C6 symmetry of the complex, a likely result given the interface structure known for this assembly [53]. The resulting IM-MS restrained ensemble is homogenous and gives rise exclusively to highly accurate models (Figure 5b). This monomodal ensemble of structures is characterized by a significantly narrower distribution of RMSD values compared with the distributions observed in Figure 5a, and is centered at an RMSD of 6 Å relative to the reference. Such RMSD values are typically achieved by our modeling workflow for structures where additional symmetry restraints can be coupled to the distances mined from IM-MS data.

For our second example, we sought to apply our method to a large, asymmetric protein complex that has been interrogated using MS methods previously [54]. The actin-related protein 2/3 (ARP2/3) complex structure was recently solved by X-ray crystallography (PDB ID 1K8K) [55]. In addition, a previous Native MS study identified two modules within the heptameric complex, the trimeric Actin Localization Module (ALM) and the tetrameric Nucleating Module (NM). Extrapolating from the data shown in Figure 3, we predict that the heptameric ARP2/3 requires between 16 and 19 internal CCS restraints to reach a PPV value of 80%. When modeling the ALM and NM individually, we find that even minimal simulated IM-MS restraint sets lead to highly accurate models. We generated high-confidence models for the trimeric ALM using two IM-derived distance restraints and a global CCS restraint. In parallel, the correct structure was readily found for the NM using 4 IM-derived distance restraints plus the global CCS restraint. These results agree well with data shown in Figure 3 for trimeric and tetrameric protein complexes.

Next, we attempted to find the minimal IM-MS restraint sets necessary for localization of ALM binding to NM, leading to a precise assignment of ALM-NM topology. We started by attempting to model this complex without providing any information about points of connectivity between ALM and NM, and filtered the resulting ensemble based on global CCS alone (Figure 6a). The resulting ensemble features two structural families, a larger population family with an RMSD distribution centered on 15 Å from the reference structure, and a less populated cluster with a very broad RMSD distribution centered on 28 Å. Interestingly, although the resolution within both families is poor, the major family appears to correctly localize the general ALM binding site on the NM surface. To reduce the ambiguity in the models, we then added two restraints that enforced connectivity between the p20 subunit of the ALM and the p34 and arp3 subunits of the NM (Figure 6b). This new connectivity information, along with the global CCS restraint, gives rise to a new ensemble of potential structures. The new restraint set acts to eliminate the majority of incorrect structures found in Figure 6a; however, it gives rise to a new, more highly resolved distribution of structures centered on 25 Å from the reference.

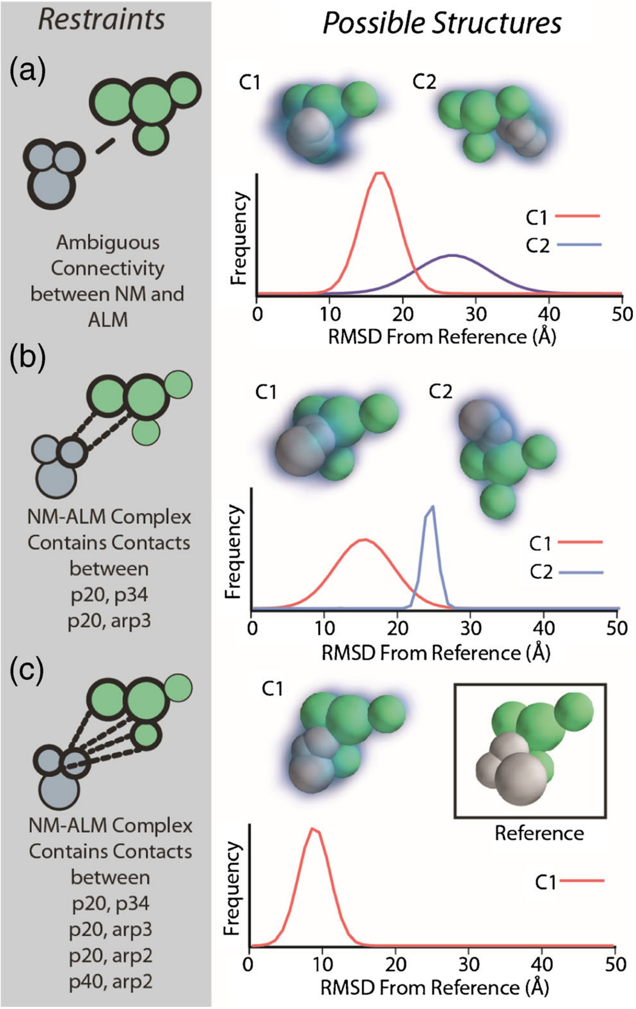


Figure 6. Docking modules within the ARP2/3 complex using connectivity restraints. After encoding the structures of the nucleating module (NM) and the actin localization module (ALM), we tested the global CCS in conjunction with various sets of connectivity restraints (left panels), (**a**), (**b**), and (**c**) for their ability to restrain the docking location and pose of NM on ALM. Structural ensembles were evaluated by hierarchical clustering and the structural families, kernel density functions, and RMSD distributions from the reference are provided

Interestingly, we note that the major structural family identified for this restraint set remains essentially unchanged from the one identified in Figure 6a, where the correct localization of ALM is determined, but having a broad RMSD distribution centered on ~15 Å from the reference structure

Finally, we applied a new restraint set with four total connectivity restraints linking p20 from the ALM with p34, arp2, and arp3 from the NM, and linking p40 from the ALM with arp2 from the NM (Figure 6c). These restraints represent the full complement of protein connectivity information accessible through MS methods [56]. When combined with sufficient connectivity information, we find that the global CCS restraint can define not only the location of ALM on the surface of NM, but also the relative orientation of the two sub-complexes. We observe a single, well-resolved family of structures centered around and RMSD value of 9 Å relative to the reference structure. Furthermore, when structures within this family are averaged, the resulting mean structure has an RMSD of only 2 Å from the reference, indicating that in this example, the mean structure is in much closer agreement with the reference than any individual structure in the ensemble. Although the mean structure in this case results in a highly accurate candidate model, we advise caution when averaging structural ensembles generally, as poorly defined structural families may cause the average structure to be distorted, or heavily biased. Combining the connectivity restraints used here with the distance and internal CCS restraints used to build models for each module, we recapitulated the correct topology using only 11 internal restraints, one-third fewer internal restraints than that the number of restraints one would predict based on PPV alone (extrapolated from Figure 3).

Conclusions and Future Directions

In this report, we explored several questions related to the generation of CG multi-protein topology models restrained using IM-MS data. We outlined a workflow based on integrative modeling principles that allows for facile translation of IM-MS data into ensembles of putative structures for hypothesis refinement or integration with high resolution docking tools. We explored the limits of coarse-grained modeling, and demonstrated that many protein topologies found in the PDB are not amenable to coarse-graining at the subunit-level, mostly because of their intricate domain architectures. However, when sufficient data is available, domain-level coarse-graining results in significant errors in only 2% of cases.

We benchmarked our CG modeling workflow against protein topologies extracted from the PDB, exploring the ambiguity in IM-MS-derived structural ensembles as a function of the information content contained in restraint sets. Our results indicated a predictable relationship between the PPV of an ensemble, and the number of internal IM-MS restraints used to generate it. Although the estimated PPV may be used as a benchmark to predict the ambiguity within a CG modeling ensemble, in many cases it underestimates the total possible information content of the IM-MS experiment, as such an analysis does not account for the structural relationships between members of an ensemble. We found that applying hierarchical clustering yields, in many cases, highly resolved conformational families that can inform future experiments, or be reported as likely structures based on available data. Additionally, we undertook two case studies that showed that highly symmetric or modular complexes can be modeled with high fidelity using smaller numbers of internal restraints than those predicted by a PPV analysis. In these cases, we observe that in large complexes the information content of the intact CCS is maximal only when one or more substructures are fully defined.

Although the computational results presented in this Critical Insight are encouraging, there are still many challenges ahead in fully harnessing the information content available in IM-MS datasets. Our CG error analysis (Figure 2) clearly motivates the development of domain-level IM-MS models of protein quaternary structure, and a move away from CG at the intact subunit level. The development of IM-MS tools for the generation of such information on protein tertiary structure, such as collision induced unfolding (CIU) [57, 58], as well as efforts to integrate IM-MS data with other sources of experimental data sensitive to local protein structure [50, 59, 60] and computational domain assignment algorithms [61] will, therefore, become increasingly important in future IM-MS protein topology modeling efforts. Similarly, our analysis of ambiguity in IM-MS models of protein quaternary structure strongly points to the need for improved methodologies capable of detecting protein complex connectivity and symmetry. As such, the development of technologies that produce a comprehensive population of protein sub-complexes, either in the gas phase or in solution, will prove highly valuable [20, 45, 46]. Finally, the ability of our IMMS_modeler algorithm to assess, for the first time, the ambiguity present within IM-MS restrained models of protein complex structure will likely lead to a greater ability to integrate such datasets with other forms of structural restraints, derived both from MS and other forms of data. Future iterations of IMMS_modeler will incorporate the ability to build models based on custom shapes, interface directly with domain-prediction software, and utilize next-generation scoring functions that enable multi-factorial assessments of model fitness. Although not discussed in detail here, it is also clear that increases in CCS precision will drive concomitant increases in the PPV of IM-MS restraints, as decreasing the current ±3% CCS error value used in the analyses described here will surely reduce the occurrence of spurious structural families within a filtered structural ensemble [62–64]. On the other hand, our data demonstrate that much can be accomplished using current IM-MS capabilities and that the proper application of restraints can be used to build high-confidence models of multiprotein complexes with both full knowledge of their precisions and informed estimates of their accuracies.

Acknowledgments

Protein topology modeling efforts in the Ruotolo lab are supported through the National Institute of General Medical Sciences, National Institutes of Health (R01 GM095832). Additionally, the authors gratefully acknowledge the support of Erik Marklund (Uppsala), Matteo Degiacomi (Oxford), and Justin Benesch (Oxford), who helped the authors to integrate IM-PACT CCS calculations into their computational workflows.

References

1. Robinson, C.V., Sali, A., Baumeister, W.: The molecular sociology of the cell. Nature **450**(7172), 973–982 (2007)
2. Marsh, J.A., Teichmann, S.A.: Structure, dynamics, assembly, and evolution of protein complexes. Annu. Rev. of Biochem. **84**, 551–575 (2015)
3. Hansen, M.R., Graf, R., Spiess, H.W.: Solid-state NMR in macromolecular systems: insights on how molecular entities move. Acc. Chem. Res. **46**(9), 1996–2007 (2013)
4. Skiniotis, G., Southworth, D.R.: Single-particle cryo-electron microscopy of macromolecular complexes. Microscopy **65**(1), 9–22 (2016)
5. Mertens, H.D.T., Svergun, D.I.: Structural characterization of proteins and complexes using small-angle X-ray solution scattering. J. Struct. Biol. **172**(1), 128–141 (2010)
6. Gingras, A.-C., Gstaiger, M., Raught, B., Aebersold, R.: Analysis of protein complexes using mass spectrometry. Nat. Rev. Mol. Cell. Biol. **8**(8), 645–654 (2007)
7. Alber, F., Dokudovskaya, S., Veenhoff, L.M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B.T., Rout, M.P., Sali, A.: Determining the architectures of macromolecular assemblies. Nature **450**(7170), 683–694 (2007)
8. Russel, D., Lasker, K., Webb, B., Velázquez-Muriel, J., Tjioe, E., Schneidman-Duhovny, D., Peterson, B., Sali, A.: Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. PLoS Biol. **10**(1) (2012). doi:10.1371/journal.pbio.1001244
9. Shi, Y., Fernandez-Martinez, J., Tjioe, E., Pellarin, R., Kim, S.J., Williams, R., Schneidman-Duhovny, D., Sali, A., Rout, M.P., Chait, B.T.: Structural characterization by cross-linking reveals the detailed architecture of a coatomer-related heptameric module from the nuclear pore complex. Mol. Cell. Proteom. **13**(11), 2927–2943 (2014)
10. Stengel, F., Aebersold, R., Robinson, C.V.: Joining forces: integrating proteomics and cross-linking with the mass spectrometry of intact complexes. Mol. Cell. Proteom. **11**(3), 1-13 (2012). doi:10.1074/mcp.R111.014027
11. Aebersold, R., Mann, M.: Mass-spectrometric exploration of proteome structure and function. Nature **537**(7620), 347–355 (2016)
12. Hyung, S.J., Ruotolo, B.T.: Integrating mass spectrometry of intact protein complexes into structural proteomics. Proteomics **12**(10), 1547–1564 (2012)
13. Smits, A.H., Vermeulen, M.: Characterizing proteins and protein interactions using mass spectrometry: challenges and opportunities. Trends Biotechnol **34**(10), 825–834 (2016)
14. Zhong, Y., Hyung, S.-J., Ruotolo, B.T.: Ion mobility-mass spectrometry for structural proteomics. Expert Rev. Proteom. **9**(1), 47–58 (2012)
15. Ruotolo, B.T., Benesch, J.L., Sandercock, A.M., Hyung, S.-J., Robinson, C.V.: Ion mobility–mass spectrometry analysis of large protein complexes. Nat. Protoc. **3**(7), 1139–1152 (2008)
16. Chorev, D.S., Ben-Nissan, G., Sharon, M.: Exposing the subunit diversity and modularity of protein complexes by structural mass spectrometry approaches. Proteomics **15**(16), 2777–2791 (2015)
17. Marcoux, J., Cianferani, S.: Towards integrative structural mass spectrometry: benefits from hybrid approaches. Methods **89**, 4–12 (2015)
18. Politis, A., Borysik, A.J.: Assembling the pieces of macromolecular complexes: Hybrid structural biology approaches. Proteomics **15**(16), 2792–2803 (2015)
19. Snijder, J., Heck, A.J.R.: Analytical approaches for size and mass analysis of large protein assemblies. Annu. Rev. Anal. Chem. **7**(1), 43–64 (2014)
20. Zhong, Y., Feng, J., Ruotolo, B.T.: Robotically-assisted titration coupled to ion mobility-mass spectrometry reveals the interface structures and analysis parameters critical for multiprotein topology mapping. Anal. Chem. **85**(23), 11360–11368 (2013)
21. Marsh, J.A., Hernández, H., Hall, Z., Ahnert, S.E., Perica, T., Robinson, C.V., Teichmann, S.A.: Protein complexes are under evolutionary selection to assemble via ordered pathways. Cell **153**(2), 461-470 (2013)
22. Benesch, J.L.P.: Collisional activation of protein complexes: picking up the pieces. J. Am. Soc. Mass Spectrom. **20**(3), 341–348 (2009)
23. Uetrecht, C., Barbu, I.M., Shoemaker, G.K., van Duijn, E., Heck, A.J.: Interrogating viral capsid assembly with ion mobility-mass spectrometry. Nat. Chem. **3**(2), 126–132 (2011)
24. Bush, M.F., Hall, Z., Giles, K., Hoyes, J., Robinson, C.V., Ruotolo, B.T.: Collision cross-sections of proteins and their complexes: a calibration framework and database for gas-phase structural biology. Anal. Chem. **82**(22), 9557–9565 (2010)
25. Bleiholder, C., Contreras, S., Bowers, M.T.: A novel projection approximation algorithm for the fast and accurate computation of molecular collision cross sections (IV). Application to polypeptides. Int. J. Mass Spectrom. **354**, 275–280 (2013)

26. Marklund, E.G., Degiacomi, M.T., Robinson, C.V., Baldwin, A.J., Benesch, J.L.: Collision cross-sections for structural proteomics. Structure **23**(4), 791–799 (2015)
27. Larriba, C., Hogan Jr., C.J.: Ion mobilities in diatomic gases: measurement versus prediction with non-specular scattering models. J. Phys. Chem. A **117**(19), 3887–3901 (2013)
28. Silveira, J.A., Fort, K.L., Kim, D., Servage, K.A., Pierson, N.A., Clemmer, D.E., Russell, D.H.: From solution to the gas phase: stepwise dehydration and kinetic trapping of Substance P reveals the origin of peptide conformations. J. Am. Chem. Soc. **135**(51), 19147–19153 (2013)
29. Shi, H.L., Pierson, N.A., Valentine, S.J., Clemmer, D.E.: Conformation types of ubiquitin M + 8H (8+) ions from water:methanol solutions: evidence for the N and A states in aqueous solution. J. Phys. Chem. B **116**(10), 3344–3352 (2012)
30. Pukala, T.L., Ruotolo, B.T., Zhou, M., Politis, A., Stefanescu, R., Leary, J.A., Robinson, C.V.: Subunit architecture of multiprotein assemblies determined using restraints from gas-phase measurements. Structure **17**(9), 1235–1243 (2009)
31. Politis, A., Park, A., Hyung, S.-J., Barsky, D., Ruotolo, B.T., Robinson, C.V.: Integrating ion mobility mass spectrometry with molecular modelling to determine the architecture of multiprotein complexes. PLoS ONE **5**(8) (2010). doi:10.1371/journal.pone.0012080
32. Hall, Z., Politis, A., Robinson, C.V.: Structural modeling of heteromeric protein complexes from disassembly pathways and ion mobility-mass spectrometry. Structure **20**(9), 1596–1609 (2012)
33. Ruotolo, B.T., Giles, K., Campuzano, I., Sandercock, A.M., Bateman, R.H., Robinson, C.V.: Evidence for macromolecular protein rings in the absence of bulk water. Science **310**(5754), 1658–1661 (2005)
34. Politis, A., Park, A., Hall, Z., Ruotolo, B.T., Robinson, C.V.: Integrative modeling coupled with ion mobility mass spectrometry reveals structural features of the clamp loader in complex with single-stranded DNA binding protein. J. Mol. Biol. **425**(23), 4790-4801 (2013)
35. Politis, A., Schmidt, C., Tjioe, E., Sandercock, A.M., Lasker, K., Gordiyenko, Y., Russel, D., Sali, A., Robinson, C.V.: Topological models of heteromeric protein assemblies from mass spectrometry: application to the yeast eIF3:eIF5 complex. Chem. Biol. **22**(1), 117-128 (2015). doi:10.1016/j.chembiol.2014.11.010
36. Zhou, M., Politis, A., Davies, R.B., Liko, I., Wu, K.-J., Stewart, A.G., Stock, D., Robinson, C.V.: Ion mobility-mass spectrometry of a rotary ATPase reveals ATP-induced reduction in conformational flexibility. Nat. Chem. **6**(3), 208–215 (2014)
37. Song, Y., Nelp, M.T., Bandarian, V., Wysocki, V.H.: Refining the structural model of a heterohexameric protein complex: surface induced dissociation and ion mobility provide key connectivity and topology information. ACS Cent. Sci. **1**(9), 477–487 (2015)
38. Schneidman-Duhovny, D., Pellarin, R., Sali, A.: Uncertainty in integrative structural modeling. Curr. Opin. Struct. Biol. **28**, 96–104 (2014)
39. Han, L., Hyung, S.-J., Mayers, J.J., Ruotolo, B.T.: Bound anions differentially stabilize multiprotein complexes in the absence of bulk solvent. J. Am. Chem. Soc. **133**(29), 11358–11367 (2011)
40. Pagel, K., Natan, E., Hall, Z., Fersht, A.R., Robinson, C.V.: Intrinsically disordered p53 and its complexes populate compact conformations in the gas phase. Angew. Chem. Int. Ed. **52**(1), 361–365 (2013)
41. Pacholarz, K.J., Porrini, M., Garlish, R.A., Burnley, R.J., Taylor, R.J., Henry, A.J., Barran, P.E.: Dynamics of intact immunoglobulin G explored by drift-tube ion-mobility mass spectrometry and molecular modeling. Angew. Chem. Int. Ed. **53**(30), 7765–7769 (2014)
42. Levy, E.D., Pereira-Leal, J.B., Chothia, C., Teichmann, S.A.: 3D complex: a structural classification of protein complexes. PLoS Comput. Biol. **2**(11), e155 (2006)
43. Arthur, D., Vassilvitskii, S.: k-means++: the advantages of careful seeding. Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms, New Orleans, LA (2007)
44. Jones, E., Oliphant, T., Peterson, P., et al.: SciPy: Open source scientific tools for Python. http://www.scipy.org/ (2001). Accessed 1 April 2017
45. Zhou, M.W., Wysocki, V.H.: Surface induced dissociation: dissecting noncovalent protein complexes in the gas phase. Acc. Chem. Res. **47**(4), 1010–1018 (2014)
46. Samulak, B.M., Niu, S., Andrews, P.C., Ruotolo, B.T.: Ion mobility-mass spectrometry analysis of cross-linked intact multiprotein complexes: enhanced gas-phase stabilities and altered dissociation pathways. Anal. Chem. **88**(10), 5290–5298 (2016)
47. Ballester, P.J., Richards, W.G.: Ultrafast shape recognition for similarity search in molecular databases. Proc. R. Soc. A **463**(2081), 1307–1321 (2007)
48. Kloczkowski, A., Jernigan, R.L., Wu, Z., Song, G., Yang, L., Kolinski, A., Pokarowski, P.: Distance matrix-based approach to protein structure prediction. J. Struct. Funct. Genomics **10**(1), 67–81 (2009)
49. Lanucara, F., Holman, S.W., Gray, C.J., Eyers, C.E.: The power of ion mobility-mass spectrometry for structural characterization and the study of conformational dynamics. Nat. Chem. **6**(4), 281–294 (2014)
50. Politis, A., Stengel, F., Hall, Z., Hernandez, H., Leitner, A., Walzthoeni, T., Robinson, C.V., Aebersold, R.: A mass spectrometry-based hybrid method for structural modeling of protein complexes. Nat. Methods **11**(4), 403–406 (2014)
51. Weiss, R.: Multivariate Density Estimation: theory, practice, and visualization. J. Am. Stat. Assoc. **89**, 359 (1994)
52. Ramachandran, P., Varoquaux, G.: Mayavi: 3D visualization of scientific data. Comp. Sci. Engin. **13**(2), 40–51 (2011)
53. Lilyestrom, W., Klein, M.G., Zhang, R., Joachimiak, A., Chen, X.S.: Crystal structure of SV40 large T-antigen bound to p53: interplay between a viral oncoprotein and a cellular tumor suppressor. Genes Dev. **20**(17), 2373–2382 (2006)
54. Chorev, D.S., Moscovitz, O., Geiger, B., Sharon, M.: Regulation of focal adhesion formation by a vinculin-Arp2/3 hybrid complex. Nat. Commun. **5**, 3758 (2014)
55. Robinson, R.C., Turbedsky, K., Kaiser, D.A., Marchand, J.-B., Higgs, H.N., Choe, S., Pollard, T.D.: Crystal structure of Arp2/3 complex. Science **294**(5547), 1679–1684 (2001)
56. Sinz, A., Arlt, C., Chorev, D., Sharon, M.: Chemical cross-linking and native mass spectrometry: a fruitful combination for structural biology. Protein Sci. **24**(8), 1193–1209 (2015)
57. Zhong, Y., Han, L., Ruotolo, B.T.: Collisional and Coulombic unfolding of gas-phase proteins: high correlation to their domain structures in solution. Angew. Chem. **126**(35), 9363–9366 (2014)
58. Eschweiler, J.D., Martini, R.M., Ruotolo, B.T.: Chemical probes and engineered constructs reveal a detailed unfolding mechanism for a solvent-free multidomain protein. J. Am. Chem. Soc. **139**(1), 534–540 (2017)
59. Hambly, D.M., Gross, M.L.: Laser flash photolysis of hydrogen peroxide to oxidize protein solvent-accessible residues on the microsecond timescale. J. Am. Soc. Mass Spectrom. **16**(12), 2057–2063 (2005)
60. Schmidt, C., Macpherson, J.A., Lau, A.M., Tan, K.W., Fraternali, F., Politis, A.: Surface Accessibility and dynamics of macromolecular assemblies probed by covalent labeling mass spectrometry and integrative modeling. Anal. Chem. **89**(3), 1459–1468 (2017)
61. Ansari, E.S., Eslahchi, C., Pezeshk, H., Sadeghi, M.: ProDomAs, protein domain assignment algorithm using center-based clustering and independent dominating set. Proteins: Struct., Funct., Bioinf. **82**(9), 1937–1946 (2014)
62. Hamid, A.M., Garimella, S.V.B., Ibrahim, Y.M., Deng, L.L., Zheng, X.Y., Webb, I.K., Anderson, G.A., Prost, S.A., Norheim, R.V., Tolmachev, A.V., Baker, E.S., Smith, R.D.: Achieving high resolution ion mobility separations using traveling waves in compact multiturn structures for lossless ion manipulations. Anal. Chem. **88**(18), 8949–8956 (2016)
63. Benigni, P., Marin, R., Molano-Arevalo, J.C., Garabedian, A., Wolff, J.J., Ridgeway, M.E., Park, M.A., Fernandez-Lima, F.: Towards the analysis of high molecular weight proteins and protein complexes using TIMS-MS. Int. J. Ion Moblity Spectrom. **19**(2/3), 95–104 (2016)
64. Glaskin, R.S., Ewing, M.A., Clemmer, D.E.: Ion Trapping for Ion Mobility Spectrometry Measurements in a Cyclical Drift Tube. Anal. Chem. **85**(15), 7003–7008 (2013)